EXPLORING USER PERCEPTION OF CAUSALITY IN AUTOMATED DATA INSIGHTS

A Dissertation
Presented to
The Academic Faculty

By

Po-Ming Law

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing
College of Computing

Georgia Institute of Technology

May 2021

EXPLORING USER PERCEPTION OF CAUSALITY IN AUTOMATED DATA INSIGHTS

Thesis committee:

Dr. Alex Endert
School of Interactive Computing
Georgia Institute of Technology

Dr. Jian Zhao
Cheriton School of Computer Science
University of Waterloo

Dr. John Stasko
School of Interactive Computing
Georgia Institute of Technology

Dr. Enrico Bertini
NYU Tandon School of Engineering
New York University

Dr. Duen Horng (Polo) Chau
School of Computational Science and Engineering
Georgia Institute of Technology

Date approved: April 9, 2021

ACKNOWLEDGMENTS

In retrospect, my PhD study was an extremely fulfilling five years. It was fulfilling not only because I have grown intellectually but because the experience made me a more confident and mature person. The kindness and generosity of the people around me played a big role in my transformation, and this small section is devoted to celebrating these people.

I would like to thank my advisors, Alex and John. It was extremely kind of you to offer to be my advisors. I am also grateful that you were willing to spend time reading my writings and giving feedback on my research. (Believe it or not, they have probably read my thesis proposal ten times!!) Having extra pairs of eyes helped me improve.

I also want to thank Rahul for being my advisor in the first few years and for being supportive and tolerant. I still remember the meeting with Rahul after I failed my qualifying exam, and your encouragement truly warmed my heart.

I would also like to thank my thesis committee for being willing to help with my dissertation. Interesting enough, I invited Enrico to join just because he was coincidentally in the same paper session as me at VIS 2020; Jian is in the committee because he had a project that was tangentially related to my research; and Polo is on board because he is the natural third vis person in the lab. I am really grateful for having all of you in my committee even though you might not see a strong connection between my research and yours.

This dissertation is made possible because of Huamin. I want to thank you for hosting me as a visiting student in Summer 2020 so I could finish a core piece of this dissertation.

I am a better person today because of a lot other people I met throughout these five years. I want to thank my mentors (Leo, Sana, Fan, and Moumita) and the buddies I met during my two internships at Adobe. I also want to give many thanks to my lab mates. (I am not going to list your names like I did in my defense because there are so many of you to thank that I know I would miss some of your names...)

I want to give special thanks to my interviewees. Some of them spare time at work to

chat with me about their visualization practice. The experience talking with them made me a more confident speaker.

Of course, I also need to thank my parents and sister. Doing a PhD in the US turns out to be one of the best decisions I made, and I am grateful that you supported me to do so.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

To facilitate data exploration and analysis, researchers have studied and developed systems that aim to automatically communicate data insights to users. For a data set of cars, these systems may proactively recommend a bar chart that depicts interesting data relationships and generate a description of the chart such as "US cars have a higher average horsepower than Japanese cars." Such automated insight functionality has emerged in commercial visualization platforms, enabling thousands of analysts to tap into its power.

Despite the wide deployment of these automated insight systems, we lack an understanding of their side effects on users during data analysis. The use of these systems is concerning because automated analysis could be unreliable: These systems may generate questionable claims about the data due to poor data quality, the violation of model assumptions, causal inference from observational data, a lack of domain knowledge, and sampling variability. The unreliability could lead to the potential misinterpretation of data and other associated consequences (e.g., poor decisions and financial losses).

This dissertation intends to advance our knowledge about the misleading side effect of automated insight systems. First, I define automated insights by reviewing the prevailing definitions of insight and highlight the criticality of creating automated insight systems with ethical considerations in mind. To understand the landscape of automated insight systems, I conducted a literature survey, identified the types of statistical information these systems provide, and summarized the sources of unreliability when this statistical information is automatically generated.

With an understanding of what are automated insights and why they could be unreliable, I interviewed 23 professional users of visualization systems to learn about their concerns with the use of automated insight systems. Some interviewees were worried about being misled by these systems. Motivated by this finding, I study whether automated insights could mislead users in reality and if they could, how to promote a correct interpretation.

Using automated insights about causation as a case study, I conducted crowdsourced studies with more than 400 participants to investigate the scenarios when misinterpretation occurs and the effectiveness of warning in preventing misinterpretation.

More broadly, this dissertation pertains to the conceptualization of human-centered artificial intelligence in data analysis systems. Throughout the dissertation, I highlight the potential ethical consequences as system developers attempt to automate aspects of data analysis. My dissertation provides empirical evidence from interviews and controlled experiments to illustrate that when automating data analysis, harmful consequences such as data misinterpretation could happen. Based on the empirical findings, it offers guidance on designing more usable and safer artificial intelligence in data analysis systems.

CHAPTER 1
INTRODUCTION

While researchers proclaim insight as a core purpose of visualization [18], they have developed automated systems that aim to communicate data insights to users. These systems do so often by recommending noteworthy visualizations and textual descriptions of data facts [78] and have been employed in various tasks during data analysis. To support data exploration, for example, automated systems could mine potentially interesting patterns and relationships in data and present them as a fact sheet (Figure 1.1). For a car data set, they may recommend a bar chart having some interesting data relationship and describe it with a statement such as "US cars have a higher average horsepower than Japanese cars." To facilitate data wrangling, these automated systems could point user attention to data quality issues that are potential concerns (Figure 1.2). They may also highlight data anomalies, missing values, and inconsistencies with the appropriate visual representation.

Underlying the development of these automated systems is often the desire to lower the barriers to and enhance the efficiency of various tasks within the data analysis workflow. For instance, by automatically surfacing noteworthy data relationships, these systems help users gain understanding of their data even when users have a very limited data analysis expertise; by highlighting potential data quality issues, they reduce users' effort to investigate anomalies in the data which is the first step in the data preparation stage.

Despite these benefits, many researchers have recognized the challenges in designing systems that automatically communicate data insights to users. First, visualization researchers have argued that insight is a sophisticated construct that embodies the characteristics of being complex, deep, qualitative, unexpected, and relevant [107]. Purely recommending potentially interesting visualizations and communicating textual descriptions of data facts are insufficient to provide data insight. This is because providing insight also

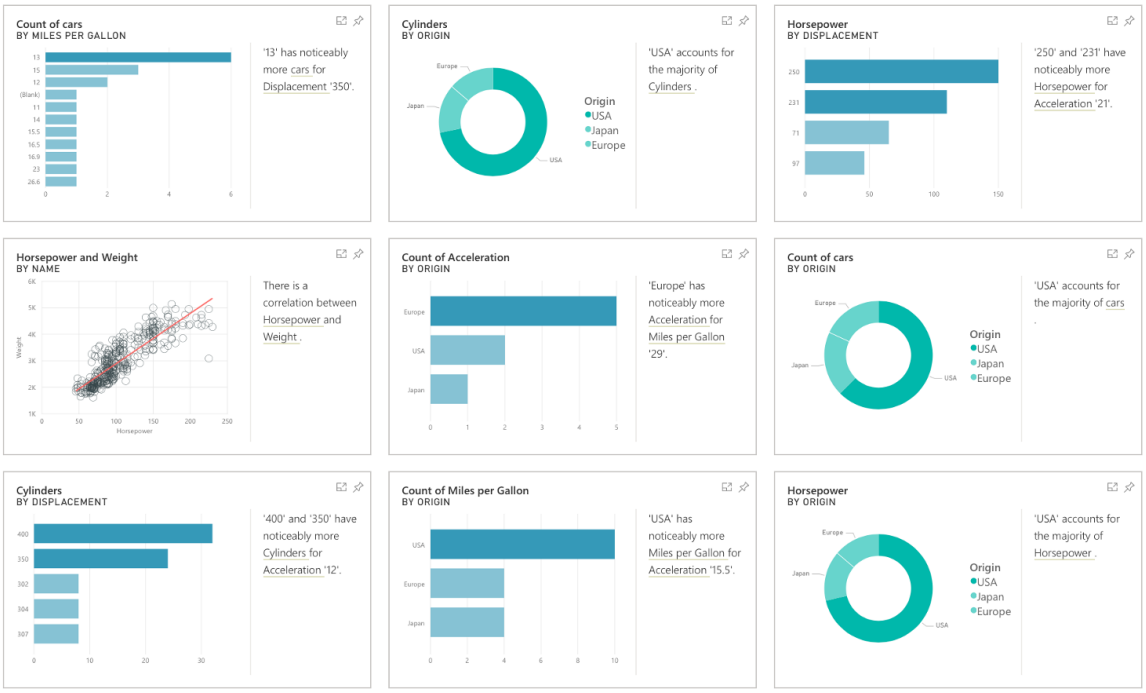


Figure 1.1: Quick Insights in Power BI [38]. Quick Insights automatically generate a fact sheet that contains noteworthy statistical information about the data.

entails endowing these computational outputs with the insight characteristics (e.g., making sure that the recommended visualizations and data facts are unexpected and relevant to users). However, how to achieve this remains an open question [79].

Second, users may easily misinterpret the visualizations and data facts recommended by these automated systems. [29]. For instance, a statement such as "US cars have a higher average horsepower than Japanese cars" only reveals statistics about a sample of data and may not be generalized to the entire population of US and Japanese cars. However, based on an observation from a data sample, people often draw an unsubstantiated conclusion about the population [29]. Creating systems that mitigate such misinterpretation presents an ongoing challenge to system designers.

In the context of investigating how to computationally generate data insights with the insight characteristics and how to do so without misleading users, some researchers use the term "automated insights" (auto-insights) to describe the computationally generated visualizations and data facts [78, 79]. My dissertation centers on studying the potential mis-

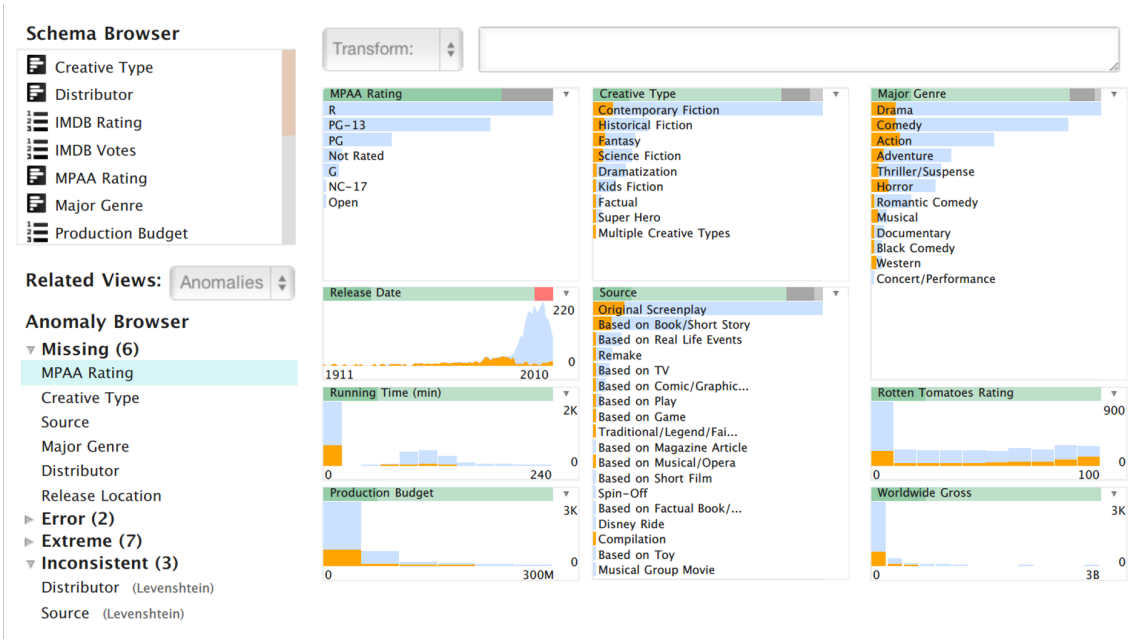


Figure 1.2: The Profile system [66]. It employs automated methods to identify data quality issues and recommends visualizations to communicate these issues.

leading power of these computational outputs, and I will employ the term "auto-insights" throughout the dissertation.

How do these auto-insights mislead users? Auto-insight systems could mislead users because they tend to be unreliable. These systems often sift through thousands of data patterns to determine noteworthy ones, sometimes leading to unreliable statistical information that would otherwise be debunked upon further investigation [29, 98]. Nevertheless, many of them are inscrutable by not providing mechanisms for validating the recommended statistical information [29, 98]. While being unreliable and inscrutable, the recommendations of visualizations and data facts possess persuasive power to shape (a potentially inaccurate) data interpretation [110]. Furthermore, by advertising themselves as ways to lower data analysis barriers, these systems naturally appeal to less-advanced users [78]. A potential lack of data analysis expertise unfortunately implies susceptibility to being misled by auto-insights. Many researchers are worried that unreliability, inscrutability, persuasiveness, and appeal to less-advanced users render auto-insight systems misleading.

Researchers' concern that auto-insight systems might mislead users is indeed shared by

visualization practitioners. In the course of developing this dissertation, I interviewed practitioners who used commercial visualization systems in their day-to-day jobs to understand their perceptions of auto-insight systems. Many of my interviewees create dashboards for a lay audience. They were worried that automation lacked context for interpreting the data and would present misleading information to their audience. I interviewed a public policy researcher who publicized data about international armed conflicts, and she commented, *"I would be worried that it [an auto-insight system] would be suggesting interpretations that are not necessarily meaningful or not make sense."*

The misleading power of auto-insights is no longer a pure concern of visualization researchers and practitioners as commercial visualization systems start to employ functionality to communicate auto-insights. Tableau alone has thousands of users worldwide [142]. The deployment of its auto-insight functionality (Figure 1.3) allows the huge user base to tap into the power of auto-insights although we still lack an understanding of the potential for auto-insights to mislead users. I argue that it is crucial to study whether auto-insights could mislead users, and if they could, how to protect users from being misled.

This dissertation intends to further an understanding of the potential misleading power of auto-insights. It surrounds three research questions (RQs) that are addressed by an interview study and a series of crowd-sourced studies.

My research endeavor started from an observation that the visualization research community knows little about how practitioners perceive auto-insight systems. While researchers have raised concerns about the use of these systems [29, 98], do practitioners' concerns align with researchers' concerns? ***What are visualization practitioners' potential concerns about the use of auto-insight systems? (RQ1)***

To understand their concerns, I conducted interviews with 23 professional users of commercial visualization systems. During the interviews, I presented auto-insight systems to probe interviewees' reactions. I identified five concerns from the interviews: misinterpretation, non-transparency, information overloading, loss of agency, and misguided data

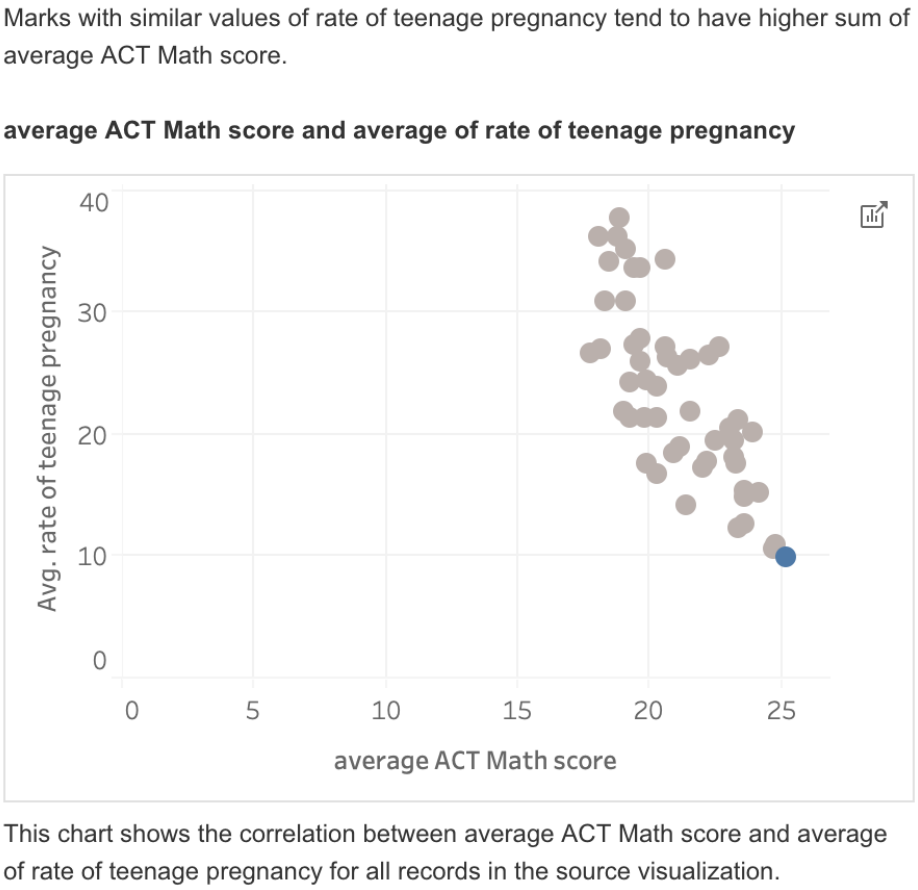


Figure 1.3: Explain Data in Tableau [145]. The user asks about the high ACT Math score in Massachusetts. Explain Data infers that teenage pregnancy rate and ACT Math score are negatively correlated, and the low teenage pregnancy rate in Massachusetts might cause the high ACT Math score.

exploration. By taking into account their experience of using visualization systems in practice, interviewees could bring a different perspective to understanding the ethical concerns in developing auto-insight systems.

Given that misinterpretation is a concern to practitioners, I strove to investigate whether auto-insights could mislead users. As auto-insights that describe causal claims (e.g., low teenage pregnancy rate in Massachusetts may lead to the high average ACT Math score in Massachusetts) have emerged in commercial systems (Figure 1.3), I focused on studying auto-insight systems that provide such auto-insights.

Figure 3 illustrates an instance where users could be misled by such auto-insights. The

user observes that Massachusetts has the highest average ACT Math score among all US states and asks the system to provide explanations for the high score. The system infers that the rate of teenage pregnancy is negatively correlated with ACT Math score and that the low rate of teenage pregnancy in Massachusetts may lead to the high ACT Math score. The auto-insight is unreliable because the system makes causal inference from observational data (as opposed to randomized experiments). The unreliability raises new concerns: ***Do people misinterpret auto-insights that describe causal claims? (RQ2) If they do, does providing a simple warning message help ensure a correct interpretation? (RQ3)***

I conducted a series of crowd-sourced studies to address these questions. In the studies, participants reviewed a series of auto-insights that described causal claims. I designed different ways to present these auto-insights and measured the degree to which participants agreed or disagreed with the causal claims across designs. The results provide evidence that auto-insight systems could utilize the persuasive power of visualizations to engender an illusion of causality. Furthermore, simply warning participants that the auto-insights might be misleading appeared to be ineffective in ensuring a correct interpretation. This implies that avoiding misinterpretation is not straightforward. Based on these findings, I advocate that system designers could encourage users to be skeptical about auto-insights and propose design ideas for doing so.

The widespread use of visualizations for data consumption has impelled many researchers to consider the ethical implications underlying the design of visualization systems. The interview study and crowd-sourced studies presented in this dissertation contribute to the discourse by providing evidence to support a thesis:

> ***User misinterpretation constitutes a concern in the use of auto-insight systems, and in particular, auto-insight systems that provide causal explanations can mislead users to interpret correlation as causation.***

More specifically, my research contributions are:

- **Articulating the side effects of automating data analysis**: I conducted interviews with 23 visualization practitioners from 12 jobs sectors to study concerns about the use of auto-insight systems during data analysis. The findings illuminated five issues in automating data analysis: misinterpretation, non-transparency, information overloading, loss of agency, and misguided data exploration.

- **Offering evidence that auto-insights could mislead users**: I conducted a series of experiments with more than 400 crowd workers to investigate whether auto-insight systems could mislead users to draw conclusions about causation by presenting information about correlation. The studies further provided insights into the efficacy of warning messages in reducing the misleading power of auto-insight systems.

The rest of this dissertation is structured as follows:

- To provide the backdrop for studying auto-insight systems, Chapter 2 summarizes existing perspectives on insight. With these perspectives, I provide the definitions of "automated insights" that will be used throughout this dissertation.

- Chapter 3 surveys the existing landscape of auto-insight systems through a systematic review of research and commercial auto-insight systems.

- Chapter 4 describes the interview study with 23 visualization practitioners that aimed to investigate their potential concerns of auto-insight systems.

- Chapter 5 presents the crowd-sourced study to understand whether people might misinterpret auto-insights that depict causal claims and whether a warning message presented alongside a causal claim could avoid misinterpretation.

- Chapter 6 discusses the future research directions surrounding the use of auto-insight systems. I further reflect on the broader implications of this dissertation.

- Chapter 7 concludes this dissertation by revisiting my research questions, thesis statement, and research contributions.

CHAPTER 2

WHAT ARE DATA INSIGHTS?

"Insight" is an ambiguous term. The Cambridge Dictionary defines insight as "(the ability to have) a clear, deep, and sometimes sudden understanding of a complicated problem or situation" [17]. In contrast, the Oxford Dictionary regards insight as "an understanding of what something is like" [109]. While the first definition considers insight a capacity to understand something, the second depicts the moment when an understanding occurs. The different dictionary definitions reveal disagreement in comprehending what insight is even among laypeople and linguists.

In the realm of data analysis, we encounter a similar conundrum. What constitutes insight in data analysis? What does it mean to automatically generate data insights? In the midst of such ambiguity, visualization researchers have provided different perspectives to consider insight. Given the vast definitions of data insight, I argue that it is difficult, if not possible, to pin down a precise definition that everyone agrees upon—this is not the goal of this chapter. My goal instead is to consider the prevailing perspectives on data insight within the visualization community, clarify the stance I am taking, and finally, articulate an operational definition of auto-insight.

2.1 Information, Observation, Finding, Knowledge, and Fact

Since information, observation, finding, knowledge, and fact have been used extensively when describing data insight, I begin the discussion of data insight by defining these terms.

Although the definitions of some of these terms appear to be less controversial, the definitions of others are more contested. For instance, whereas Ackoff [1] defined information as "data that are processed to be useful, providing answer to who, what, where, and when questions," Min et al. [22] consider information "data that represents the results of a compu-

Table 2.1: The definitions of information, observation, finding, knowledge, fact, data fact, insight, and auto-insight. I highlight the definitions I employ in this dissertation in bold.

Term	**Definition**
Information	Information is **data about something**. (e.g., a numerical value about an entity and a textual description about an event). It is an encapsulating term that subsumes observation, finding, knowledge, and fact.
Observation	Observation is **information gleaned from data analysis**. I do not make a distinction between observation and finding.
Finding	Finding is **information gleaned from data analysis**. I do not make a distinction between observation and finding.
Knowledge	Knowledge is **information acquired by humans through cognitive processes such as perception, learning, and reasoning**. Knowledge often resides in human brains but can also be represented in other media (e.g., books and databases).
Fact	There are two prevailing definitions of fact. Fact is often defined as **information that is proven to be true** [116]. Sometimes, fact is also defined as information that can be verifiable (i.e., can be proven to be true or false) [157]. Different from the first definition, the second definition is broader and considers false information as fact. I employ the first definition in this dissertation.
Data fact	Data fact commonly refers to a textual description about statistical information obtained from the data [136] (e.g., "US cars have a higher average horsepower than Japanese cars" for a data set of cars). Throughout this dissertation, I argue that such a textual description about the data could be inaccurate and may not be true (see Section 3.4 in particular). Hence, data facts may not be facts if we restrict fact only to refer to information that is proven to be true (the first definition above). However, data facts are facts if we define fact broadly as information that can be verifiable (the second definition above). In summary, I consider data fact as **statistical information about the data in the form of a textual description regardless of its truthfulness**.
Data insight / Insight	Data insight is **a unit of knowledge from data that possesses the insight characteristics** (e.g., complex, deep, qualitative, unexpected, and relevant [107]).
Automated data insight / Auto-insight	I define auto-insight as **statistical information in the form of visualization and/or data fact recommendations**. Note that the degree to which people consider that information to be insightful could vary.

tational process, such as statistical analysis, for assigning meanings to the data." Whereas Bertini and Lalanne [9] regard knowledge as "justified true belief," Liew [88] commented

that knowledge is "the cognition or recognition, capacity to act, and understanding that reside or is contained within the mind or in the brain."

Table 2.1 summarizes my definitions of information, observation, finding, knowledge, fact, and other related terms. Instead of adding fuel to the philosophical debate, my objective here is to provide loose operational definitions that are relatively well-accepted. My definitions could serve as the start of a conversation about the use of these terms, and I expect the definitions to shift as researchers engage in debates to clarify their meanings.

In this dissertation, *information* refers to data about something (e.g., a numerical value about an entity and a textual description about an event). It is an encapsulating term that subsumes observation, finding, knowledge, and fact. *Observation* and *finding* are information gleaned from data analysis. I do not make a distinction between these two terms. *Knowledge* is information acquired by humans through cognitive processes such as perception, learning, and reasoning. Knowledge often resides in human brains but can also be represented in other media (e.g., books and databases).

Fact has two common definitions. Fact is often defined as information that is proven to be true [116]. Sometimes, fact is also defined as information that can be verifiable (i.e., can be proven to be true or false) [157]. Different from the first definition, the second definition is broader and considers false information as fact. I employ the first definition in this dissertation because it appears to be more commonly used among laypeople.

A term related to fact is *data fact*. It commonly refers to a textual description about statistical information obtained from the data [136] (e.g., "US cars have a higher average horsepower than Japanese cars" for a data set of cars). Throughout this dissertation, I argue that such a textual description about the data could be inaccurate and may not be true (see Section 3.4 in particular). Hence, data facts may not be facts if we restrict fact only to mean information that is proven to be true (the first definition of fact above). However, data facts are facts if we define fact broadly as information that can be verifiable (the second definition of fact above). In summary, I consider data fact as information about the data in the form

of a textual description regardless of its truthfulness.

2.2 Prevailing Perspectives on Data Insight

With the nomenclature, I now turn to discussing the prevailing perspectives on data insight. Broadly, visualization researchers have considered insight from three perspectives: a unit of information or knowledge, a psychological state, and an analysis by-product.

The visualization community often defines insight as a unit of information or knowledge [126]. A widely accepted definition is offered by North [107] who describes insights as complex, deep, qualitative, unexpected, and relevant revelations. Chang et al. [21] observed that visualization researchers often use "insight" in the same sense as "knowledge" or "information." Chen et al. [23] define insight as a fact that is evaluated through a mental model to inspire a psychological state of enlightenment. Researchers who employ insight-based evaluation methods [115] for visualization evaluation often define data insights as facts, observations, generalizations, and hypotheses from data (e.g., [47, 92, 169]). In characterizing the insights quantified selfers gained from their data, Choe et al. [25] regarded data observations such as trends and comparisons as insights.

Insight can also be viewed as a psychological state. Laypeople often describe insight as a eureka moment [8], a light bulb moment [148], and an aha moment [19]. In cognitive science, a prevailing view is that insight occurs when people transition from a state of not knowing how to solve a problem to a state of knowing how to solve it [95]. Chang et al. [21] called this type of insight a spontaneous insight and noted that insights as units of knowledge serve a knowledge-building function that promotes spontaneous insights.

From the first perspective, insight is considered a finding—the end result of an analysis. A contrasting view is to regard insight as a by-product of data analysis. Yi et al. [168] believe that insights can be "sources or stimuli of other insights" (e.g., insightful questions that an analyst had not thought about). Stasko [137] commented that the effects of knowledge on an analyst's mental model (e.g., learning a domain and confirming a hypothesis)

are also insights.

2.3 Data Insights and Knowledge

Throughout this dissertation, I consider data insight from the first perspective (data insight as a unit of knowledge from data) for its prevalence in the visualization research community. I note, however, that visualization researchers often do not view data insight and knowledge equivalent. Instead, many researchers deem insight as "non-trivial" knowledge [115] that could potentially be "complex, deep, qualitative, unexpected, and relevant" [107]—insight is knowledge with certain characteristics that could be difficult to articulate.

Some characteristics of insight that recur in the literature include relevant [107], trustworthy [79, 124], and complex [107]. A piece of knowledge is relevant when it is tied to the outcome an analyst seeks through data analysis. For example, in business enterprises, the goal of data analysis is often to drive better decisions [67, 79]. Relevance therefore depends on the ability of a piece of knowledge to inform actions. For users to consider data observations insightful, they should also develop a sufficient level of trust in the observations [124]. Without sufficient trust, users would likely refute the observations instead of considering them data insights. Also, reports of data insights often depict the assembly of multiple information sources (e.g., domain knowledge and context) to help construe data findings in a new light [79]. This construction of insight using data findings, domain knowledge, and context explains its complexity.

Besides relevance, trustworthiness, and complexity, other researchers believe that data insights are spontaneous (characterized by an aha or eureka moment) [21, 79], actionable (inform decision making) [79], unexpected (diverge from one's expectations) [79, 107], confirmatory (confirm one's expectations) [79], and personal (echo with one's educational background, political affiliation, and personal experience) [112]. This set of diverse characteristics reveals an ongoing challenge to distinguish between data insights and knowledge.

2.4 Capturing and Measuring Insight Characteristics

Researchers' endeavor to investigate insight is not limited to compiling a list of characteristics that make a piece of knowledge insightful. They have also striven to develop methods for recording data insights and measuring some of the insight characteristics.

Capturing and measuring the insight characteristics entails developing methods for recording data insights. Supporting insight provenance is challenging because insight is not directly observable [118]. Recording insights often requires users to annotate the data observations (e.g., [114]) or verbalize the thought process (e.g., [119]). Yet, these approaches are cognitively demanding and can interrupt analytic reasoning [39]. To minimize cognitive effort from users, analytic provenance researchers have studied recording insights through eye movements (e.g., [64]) and interaction logs (e.g., [13]) that are less intrusive. In particular, Gotz and Zhou [48] proposed to organize interaction logs into semantically meaningful actions (e.g., query and filter) for recording the process through which insight is obtained. Dou et al. [39] further provided evidence that a significant part of the insight generation process could be recovered from interaction logs.

With records of insights, investigators can quantify the characteristics of an insight (e.g., how unexpected a finding is) by qualitative coding. In insight-based evaluation, participants are asked to perform open-ended analysis tasks using a visualization and report the findings in a think-aloud manner [126]. Domain experts then assign numerical scores (e.g., unexpectedness) to the findings to measure the insight characteristics [126]. However, the standard insight-based evaluation does not capture the relationship between interactions and insight generation. Motivated by this limitation, Guo et al. [51] augmented the standard approach by logging user interactions during insight-based evaluation and conducting correlational analysis to identify the interactions or system features that lead to certain insight characteristics. For example, their approach could enable investigators to look into whether a particular interaction (e.g., locating a node in a network visualization) is correlated with

unexpectedness in data insights.

2.5 Endowing Information with Insight Characteristics

Instead of capturing and measuring the insight characteristics, this dissertation concerns a different yet related problem—endowing information with the insight characteristics. As we have seen in the introduction, automated systems that aim to provide data insights often do so by recommending potentially interesting visualizations and textual descriptions of data facts (e.g., [38]). For a car data set, these systems may recommend a bar chart along with a textual description such as "US cars have a higher average horsepower than Japanese cars." Despite the goal of these systems to provide insight, some researchers feel that the information depicted by the recommendations lack the insight characteristics [138]: Contrary to complexity [107], relevance [107] and trustworthiness [79, 124] that characterize insights, the information often seems shallow, irrelevant, and untrustworthy.

This realization has prompted researchers to consider methods for conferring the insight characteristics on the recommendations. Arguably, designing systems that recommend noteworthy statistical information with all the characteristics of insight could be challenging. However, I believe that promoting some of the insight characteristics is possible. To system developers, the benefits of doing so could also be huge: These systems will see a wider adoption as users find the recommendations deep, nuanced, relevant, and trustworthy.

Hence, my collaborators and I have proposed considerations for providing recommendations that users would consider insightful [79]. For example, to inspire user *trust* in the recommendations, systems could allow users to validate the automatically generated statistical information and enhance the transparency in the generation process. To provide more *relevant* recommendations, these systems could acquire mental models (e.g., expectations about the data) from users. For instance, in concept-driven visual analytics [26], users can articulate their expectations about the data through natural language (e.g., I expect the US

to have the highest GDP per capita in the world), and the system will respond accordingly (e.g., showing a bar chart that reveals the GDP per capita of different countries with a textual description that indicates the true rank of the US).

Throughout this dissertation, I employ the framing of visualization and data fact recommendations as "auto-insight" to capture the emergent interest to endow these recommendations with the insight characteristics.

2.6 Ethical Considerations for Automating Data Insights

I argue that the discourse on auto-insight would be incomplete by not deliberating the potential consequences as the visualization and data fact recommendations acquire the insight characteristics. Visualizations and textual descriptions hold significant power to sway data interpretation. For example, data visualizations can make a message more persuasive [110]. Slanted titles can influence the perceived message of a visualization without viewers' awareness of the bias [74, 75]. Auto-insight systems could abuse such rhetorical and persuasive power as system developers endow the visualization and data fact recommendations with certain insight characteristics.

A potentially concerning insight characteristic is trustworthiness. Trust in auto-insight systems could be unwarranted because auto-insight systems are often unreliable, non-transparent, and inscrutable [29, 98]. By sifting through numerous data patterns to identify interesting ones, they may reveal noise in data rather than tenable knowledge; by not providing explanations for the recommendations, they conceal the unreliability in the recommended data observations; by not offering validation mechanisms, they preclude users from verifying the veracity of the recommendations. Unreliability, non-transparency, and inscrutability could turn trust into blind trust, thereby misleading users into believing potentially false claims about the data.

Hence, as we design methods to provide recommendations that users would find insightful, it is important to contemplate the moral character behind our designs. Such ethical

considerations provide the broader context in which this dissertation situates.

2.7 Defining Automated Data Insights

I hope that this discussion clarifies the prevailing perspectives of data insights and the definition of data insights I have adopted in researching the automatic production of data insights. Here, I offer an operational definition of auto-insight that will be used in the rest of this dissertation:

> *Automated insight is statistical information about the data in the form of visualization and/or data fact recommendations, and this information could be appropriately (and ethically) endowed with the insight characteristics as it is automatically generated.*

Defining insight has long been an ongoing debate in the visualization community [21, 107]. Instead of closing this debate, I intend to provide a starting place for considering what is auto-insight, with the expectation that the definition will be expanded and refined through future conversations among researchers. For example, in this dissertation, I only regard statistical information (e.g., correlation and causation) in the form of visualization and/or data fact recommendations as auto-insights. However, users might consider recommendations beyond statistical information (e.g., what actions to take) insightful, and future variations of the definition could include these recommendations.

In Section 2.5 and Section 2.6, I have already argued for the importance of conferring the insight characteristics on the visualization and data fact recommendations and the criticality of doing so with ethical considerations in mind. Therefore, the above definition emphasizes that auto-insights could be "appropriately (and ethically) endowed with the insight characteristics." I wish this emphasis in the definition would impel system developers to investigate techniques through which to provide more insightful statistical information to users and encourage researchers to contemplate the moral character of the algorithmic

recommendation of statistical information during data analysis.

Finally, the term "auto-insight" might carry the connotation of automating some internal cognitive processes through which an insightful moment comes about when users receive some information. I would like to clarify that this is not what I mean when I use the term—it is unlikely that we can automate the internal cognitive process through which information is being internalized. Instead, I use "auto-insight" to refer to the automatic production of statistical information and the provision of the information in ways so that users deem it insightful (e.g., by extracting information that is more relevant to users).

CHAPTER 3
CHARACTERIZING AUTOMATED DATA INSIGHTS

While the previous chapter offers a definition of auto-insight, this chapter presents examples of user interfaces that recommend auto-insights through a systematic review of these interfaces. This review intends to provide more clarity to the kind of systems I study in this dissertation.

As auto-insight systems could mislead users by recommending potentially wrong statistical information, it would be useful to look into the types of statistical information they provide. In surveying auto-insight systems, I further identified 14 types of automatically generated statistical information. One of them is causation, which is my focus when studying the misleading power of auto-insights (Chapter 5). In the following, I present the methodology of the systematic review and the 14 types of statistical information.

This chapter is an extended version of a 2020 IEEE VIS short paper I co-authored with Endert and Stasko [78].

3.1 Reviewing Relevant Systems

As a first step to collect a set of relevant systems for the review, I defined auto-insight systems. I regarded auto-insight systems as user interfaces that automatically recommend visualizations or data facts to users. I also considered proactive behaviour an important characteristic of auto-insight systems: As the very goal of auto-insight systems is to automatically communicate data insights, they should proactively identify potentially interesting data relationships and present them to users. Furthermore, I focused on systems that extract recommendations from tabular data since it is one of the most common data types [15, 16].

These considerations excluded several recommendation algorithms or systems from

the review. I omitted publications about algorithms that extract visualizations or data facts because they do not have a user interface (e.g., [147]). I also omitted recommendation systems that do not proactively identify noteworthy data relationships. For example, while Voyager [163, 164] and Show Me [94] recommend perceptually effective visualizations, they do not proactively identify prominent data relationships and are excluded from the review. Finally, I excluded some systems that automatically annotate visualizations such as Contextifier [61] and NewsViews [46] because these systems do not generate annotations directly from the data. Instead, they create annotations by associating news articles to points of interest in a visualization.

My review included both research systems and commercial products. To collect relevant research systems, I reviewed academic publications. First, I gathered a set of seed papers about systems that recommend data facts or visualizations. For systems that recommend data facts, I started with two publications that explicitly used the term "data facts" to describe the recommendations by the systems [136, 156]. For systems that recommend visualizations, Lee [81] reviewed 12 relevant research systems. I omitted two systems that do not have a user interface [5, 127]. Hence, this process yielded 2 (papers about data fact recommendation) + 10 (papers about visualization recommendation) = 12 seed papers.

I then gathered publications citing the 12 seed papers from Google Scholar and those cited by the seed papers. I found 833 unique publications. Having collected the set, I reviewed publications at ten relevant venues (InfoVis, VAST, TVCG, EuroVis, PacificVis, SIGCHI, AVI, VLDB, SIGKDD, and SIGMOD). The review resulted in ten additional papers that depict auto-insight systems [7, 14, 38, 66, 76, 77, 90, 96, 99, 131]. I repeated the process by collecting publications citing the ten papers and those cited by the ten papers and found 298 new unique publications. However, I did not find additional relevant publications from this set. The whole process yielded 12 (systems from the seed papers) + 10 (systems from the additional papers) = 22 research systems. I gathered these systems in Dec 2020.

For commercial products, I included Explain Data in Tableau [145] and Quick Insights

in Power BI [38]. As Quick Insights was already described by one of the 22 research publications, I only added Tableau Explain Data to the 22 research systems. Hence, in total, I found 23 auto-insight systems for the review. While searching for relevant commercial products, I also discovered other well-known commercial auto-insight systems (e.g., Wordsmith offered by a company called Automated Insights [6] and Quill created by Narrative Science [103]). However, these products were not available for free, and they did not have documentations that detail the auto-insight functionality. Hence, I removed these products from the review.

3.2 Coding the Types of Statistical Information

I then coded the statistical information automatically extracted by the 23 auto-insight systems. I used the fact taxonomy proposed by Chen et al. [23] as a foundation for the analysis because of its comprehensiveness. The authors curated the taxonomy based on the data observations by the users of Many-Eyes [153], a well-known collaborative visualization platform. They further grounded the taxonomy in several prior taxonomies [132, 4] to ensure the comprehensiveness. Based on this taxonomy, I reviewed the auto-insight systems and identified 14 types of statistical information that could be automatically extracted. The 14 types comprise the 12 types of facts in Chen et al.'s taxonomy and two additional types (visual motifs and causation) newly found during the review.

When reviewing the academic publications, I observed that some publications tend to be vague in describing the types of statistical information the systems support. For example, some systems provide a flexible framework that enable developers to incorporate functionality to extract additional statistical information. However, they do not explicitly state the types of statistical information that were included in the current implementation. While coding the types of statistical information, I looked for explicit statements about what types of statistical information the systems extract in the paper. I also examine the accompanying figures, videos, and systems if they were publicly available.

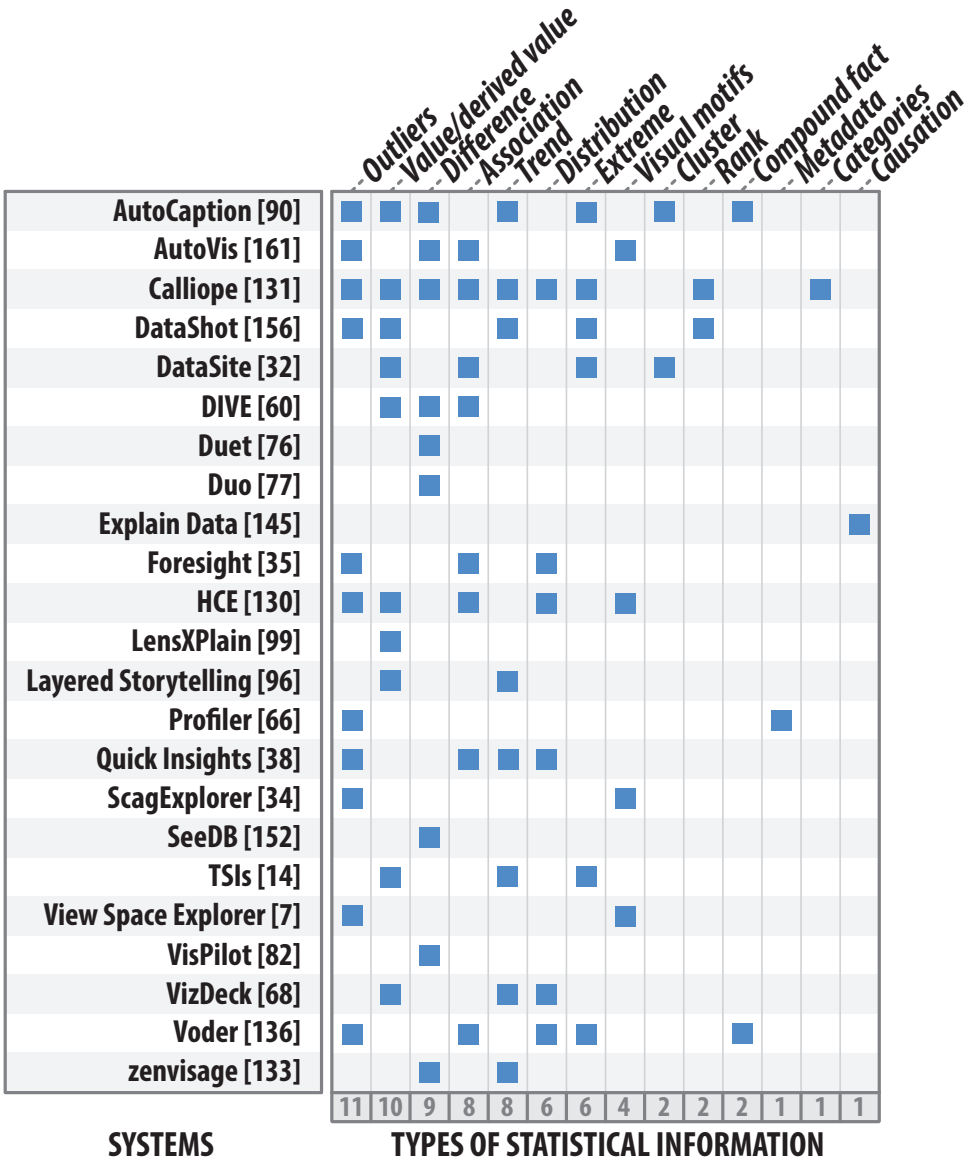


Figure 3.1: The 14 types of statistical information and their frequency of occurrence in the 23 auto-insight systems I reviewed.

3.3 14 Types of Statistical Information

During the review, I observed various approaches to extracting statistical information from data. Visualization recommendation systems often score and rank charts to prioritize the presentation of the charts. For example, Foresight (Figure 3.2) ranks scatterplots by correlation coefficient [35]. Systems that communicate data facts may present such facts based on the attributes shown in a chart or areas of interest in a visualization. For instance,

Voder (Figure 3.3) generates textual descriptions in relation to the attribute combination in a chart [136]. Temporal summary images (Figure 3.4) identify salient regions in a visualization and annotate them with descriptive facts [14]. In the following, I describe the 14 types of statistical information identified from the systematic review. I summarize the results in Figure 3.1.

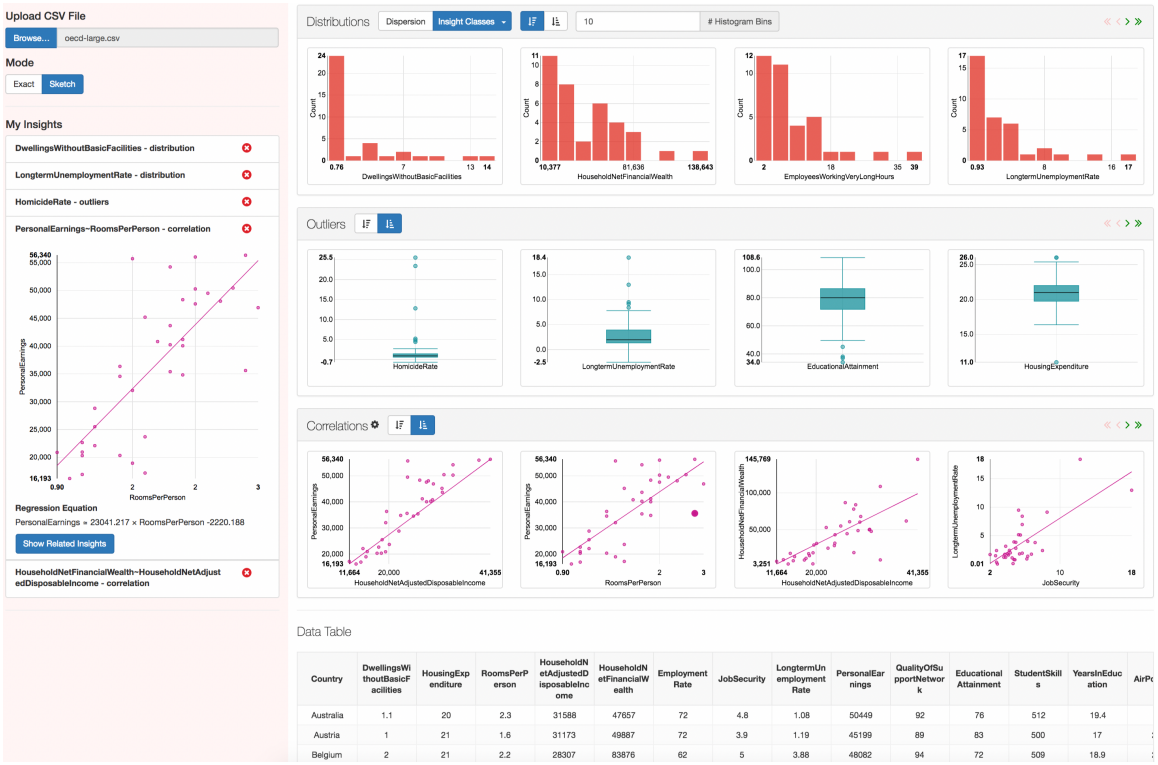


Figure 3.2: The Foresight interface [35]. Foresight ranks charts based on the statistical properties (e.g., correlation) in the data.

Outliers. Statistical information about outliers is available in 11 auto-insight systems. 6/23 systems provide statistical information about outliers in a single variable. For example, Voder categorizes data values that are 1.5 time the interquartile range below the first quartile or above the third quartile as outliers [136]. 5/23 systems provide statistical information in two dimensions. Notable examples include systems that employ scagnostics for ranking scatterplots (e.g., [7, 34]). These systems utilize the outlying scagnostics measure to identify scatterplots that have outlying data points. Finally, Quick Insights [38] and Calliope [131] communicate outliers in a time series that are highlighted in a line chart.

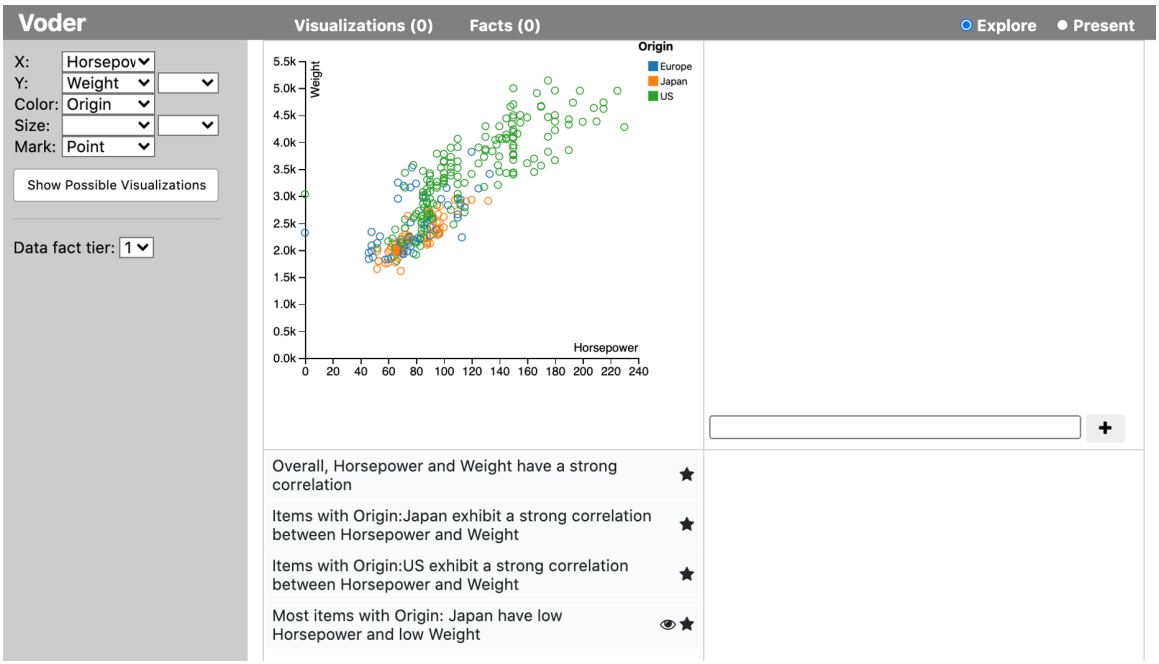


Figure 3.3: The Voder interface [136]. Voder generated textual descriptions of data facts based on the attributes shown in a chart.

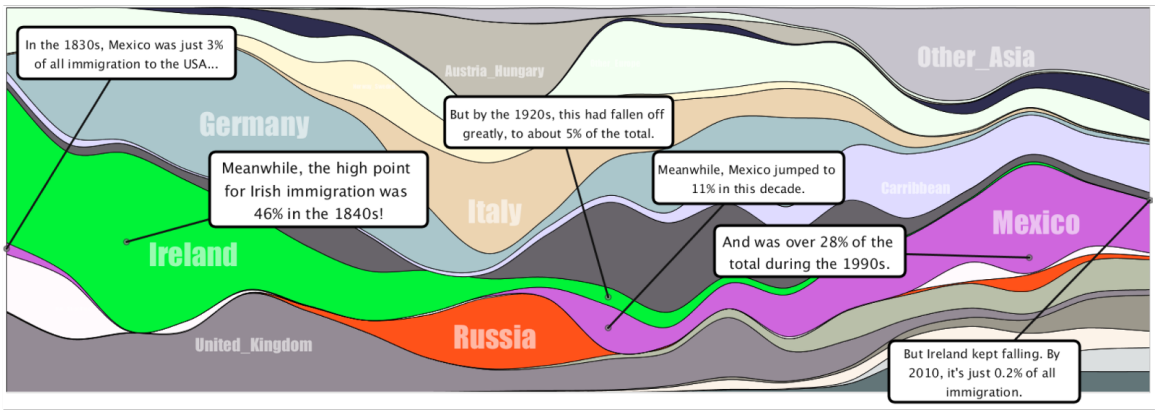


Figure 3.4: Temporal summary images automatically annotate the prominent regions in a visualization [14].

Value/derived value. Statistical information about a value in a single row in a table or a value derived from multiple rows in a table appear in 10/23 systems. The multimodal layered storytelling approach reveals prominent values in a timeline [96]. VizDeck measures the number of unique categories in a categorical variable and ranks bar charts by unique category count [68]. DataSite finds the average of a numerical variable [32] while DataShot computes the proportion of categories in a categorical variable [156]. Both systems state

the derived values (average and proportion) as textual descriptions.

Difference. Statistical information about difference involves quantitative comparison between distributions. 9/23 systems provide such statistical information. With Duo [77], users can specify two groups of objects (e.g., cities in China and cities in the US). For each attribute (e.g., population), Duo compares the two groups to determine whether they have different distributions [77]. SeeDB ranks grouped bar charts by computing the earth mover's distance between the two probability distributions depicted in the charts [152]. AutoVis conducts a one-way ANOVA for charts that show continuous-categorical data and sorts the charts by p-value [161].

Association. 8/23 systems provide statistical information about association (i.e., quantitative relationship between two numerical variables). These systems commonly identify either linear relationship using Pearson correlation (8/23) or non-linear relationship using more sophisticated measures (1/23). DataSite computes a Pearson correlation between two variables and presents it as a textual description alongside a scatterplot [32]. Hierarchical Clustering Explorer (HCE) ranks scatterplots by least-squares error curvilinear regression and quadracity to identify scatterplots that show a quadratic relationship [130].

Trend. I found statistical information about temporal trends in 8/23 systems. These systems extract upward and downward trends (5/23), steady trend (2/23), and periodicity (2/20). With the ZQL language for specifying visualizations, zenvisage can order line charts based on the upward trends they show [133]. Temporal summary images add annotations to the flat region in a time series visualization [14]. Quick Insights extracts time series that show seasonality [38].

Distribution. 6/23 systems communicate statistical information about the distribution of a variable. Voder presents data facts about the range of a numerical variable [136]. Foresight ranks charts based on several measures of distribution including dispersion, skewness, and heavy-tailedness to reveal charts with a noteworthy distribution [35].

Extreme. 6/23 systems show statistical information about the minimum and maximum

values in a stream of values. For example, DataSite [32] and Voder [136] present the minimum and maximum values in a numerical variable as textual description. Temporal summary images annotate the lowest and highest points in a time series [14].

Visual motifs. Visual motifs are unique visual patterns in a chart that do not fall into other auto-insight types. They include the special patterns in scatterplots identified by the scagnostics measures [158]. For example, the striated measure detects scatterplots with parallel lines. 4/23 systems identify visual motifs in scatterplots by utilizing scagnostics [7, 34, 161] or other measures (e.g., uniformity) [130].

Cluster. 2/23 systems recommend statistical information about clusters. For example, DataSite employs K-means and DBSCAN to find clusters in a scatterplot [32]. AutoCaption employs a deep learning approach for cluster detection and describes the clusters using natural language [90].

Rank. Such statistical information involves sorting categories by a numerical variable and appears in 2/23 systems. For a dataset of cars, DataShot recommends a data fact that says, "Compact, SUV, Midsize are the top 3 categories in the year of 2008" [156]. This information is generated by ranking different types of cars by the numerical variable sales.

Compound fact. Chen et al. [23] defined a compound fact as "a fact that contains two or more facts." Voder recommends a fact by combining statistical information about both a derived value and distribution [136]. For example, it generates the fact "Average Retail Price of SUV is 1.76 times Sedan" for a car dataset [136]. The fact includes a derived value (average) and is about the distribution of retail price.

Metadata. Statistical information about metadata concerns data about a dataset. Such information includes missing values and other data quality issues [23]. Profiler uses detection routines to identify data quality issues and suggests charts to visualize the issues [66].

Categories. Only Calliope [131] provides statistical information about categories. This type of statistical information enumerates a set of categories in the data, and these categories are not ranked. An example is "All in all, jobs in this data can be classified into 4

categories: rich, middle, lower middle, and lower" [131].

Causation. Statistical information about causation is a specific type of compound fact that describes cause and outcome. For example, when exploring a bike sharing dataset using Tableau, a user may observe from a bar chart that people ride the longest distance in August. The user can select the mark for August and asks Explain Data to provide an explanation (Figure 3.5 top). Explain Data will infer that the ride distance is the longest in August (an outcome) and suggest that the long distance is potentially due to a large number of overcast days in the month (a cause) (Figure 3.5 bottom).

3.4 Unreliability in the Recommendation of Statistical Information

This dissertation intends to investigate the misleading power of auto-insights. The above list can provide a starting point to help understand how auto-insight might mislead users. A core issue is that statistical information often involves uncertainty and unreliability stemmed from multiple sources:

Poor data quality: In data analysis, it is well-known that garbage in, garbage out—the quality of data analysis highly depends on whether analysts have good data to begin with. This principle similarly applies to automated analysis.

Violation of assumptions: Auto-insight systems employ models (e.g., regression models) and metrics (e.g., correlation) for mining prominent data relationships. For the conclusions to be valid, these models and metrics often require a set of assumptions to be satisfied. For instance, Pearson correlation assumes the absence of outliers [129]. When there are outliers in the data, auto-insight systems can produce untenable claims about correlation.

Causal inference from observational data: Statistical information about causation could be unreliable because of the inherent difficulties to make causal inferences from observational data (as opposed to randomized experiments) [123]. Auto-insight systems can easily mistake mere correlation between two events for causation. Claims about causation without warning could lead users to draw conclusions about causal connection between events

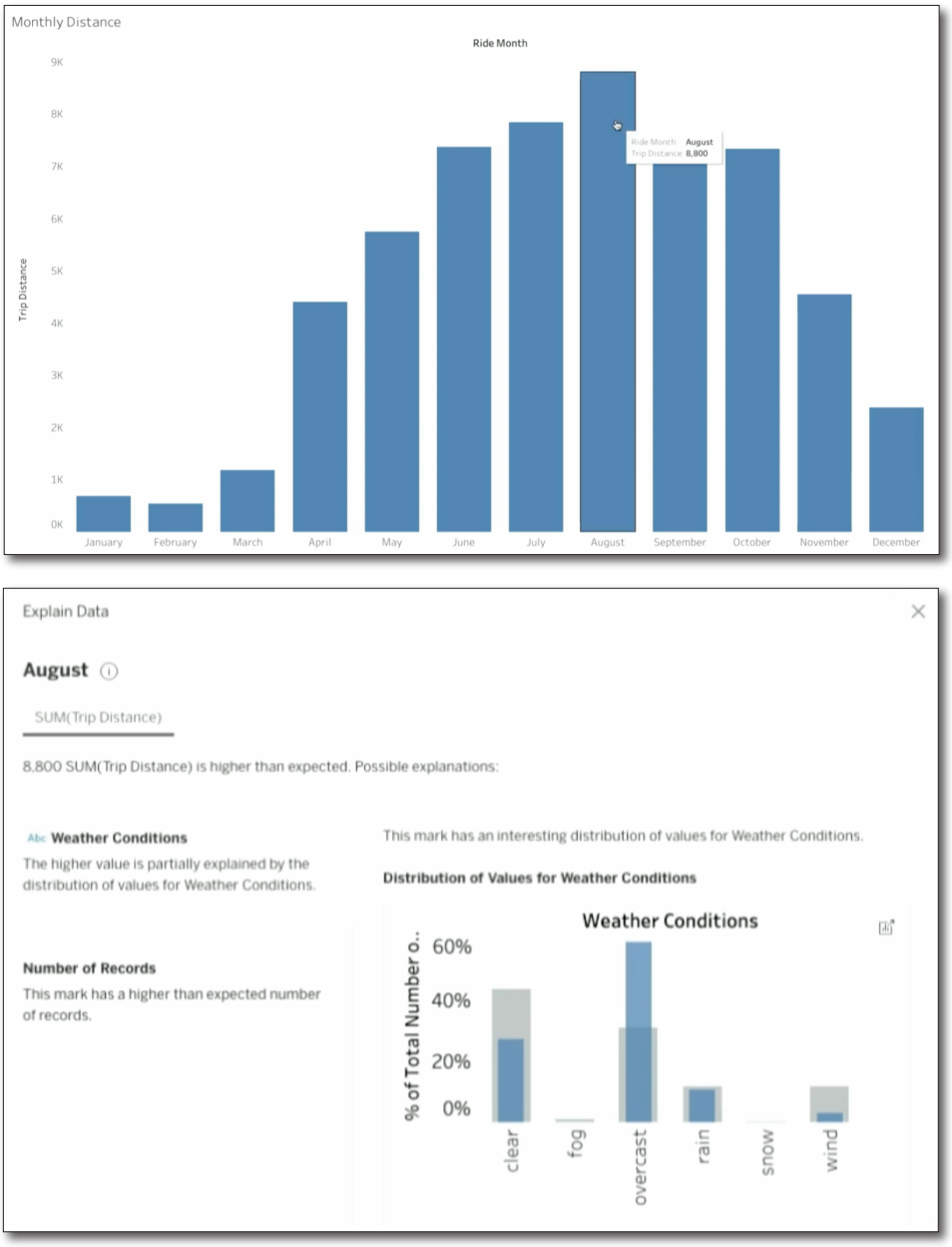


Figure 3.5: A usage scenario of Tableau Explain Data [145]. The user is exploring a bike sharing dataset and observes a long ride distance in August (top). Explain Data infers that a potential cause could be that there are many overcast days in August (bottom).

that are indeed unrelated.

Lack of domain knowledge: Auto-insight systems often lack information about the analysis context and data semantics. Without such information, these systems may fail to interpret and process the data correctly. For example, while some data attributes (e.g., monthly

sales) can be summed to find the total, others (e.g., monthly temperature) should be averaged. Not knowing the semantics of attributes can lead to data aggregation that does not make sense [105].

Sampling variability: Observing correlation in a data sample does not imply that the correlation exists in the population because of sampling variability [58]. For example, a claim about correlation between two variables in the data can be misleading because users might believe that the same correlation occurs in the real world.

In the following chapters, I will highlight the misleading power of auto-insight as a potential concern to visualization users (Chapter 4). Using statistical information about causation as a case study, I will then investigate whether auto-insights mislead users and the efficacy of warning to avoid potential misinterpretation (Chapter 5).

CHAPTER 4
UNDERSTANDING USERS' CONCERNS ABOUT AUTOMATED DATA INSIGHTS

Visualization researchers have voiced concerns about the use of auto-insight systems. For example, Correll [29] is worried that auto-insight systems can stifle user agency by providing excessive guidance. He further commented that these systems could provide false conclusions by serving as a "p-hacking machine" that enumerates data patterns without validation. Echoing Correll's concerns, McNutt et al. [98] suggested that these systems could be opaque (provide recommendations that are not interpretable), inflexible (fail to consider domain knowledge while producing recommendations), brittle (exhaustively search through data patterns without proper validation), and domineering (reduce user agency with excessive guidance).

Despite this effort in the research community to articulate the potential pitfalls of auto-insight systems, we know little about users' concerns about their use. This chapter addresses the first research question I introduced in Chapter 1: ***What are visualization practitioners' potential concerns about the use of auto-insight systems?*** A lack of considerations for users' workflow and concerns has been recognized as a likely reason for the failed adoption of intelligent systems [102]. Hence, understanding the concerns of the potential users is crucial for designing auto-insight systems that these users consider usable and are willing to adopt. Furthermore, through their practical experience in data analysis, these users can more accurately assess the potential pitfalls of auto-insight systems, thereby enriching the discussions in the research community.

Auto-insight functionality has been deployed in some research and commercial visualization systems (e.g., Tableau [145] and Power BI [38]). Target users include people who use these visualization platforms in their day-to-day work. Thus, I interviewed 23

professional users of visualization platforms (hereafter, *visualization users*) to investigate their potential concerns about using auto-insight systems. In this chapter, I first detail the methodology of the interview study (Section 4.1). I then describe the data analysis workflow of interviewees (Section 4.2) and report on the five concerns interviewees had: misinterpretation, non-transparency, information overloading, loss of agency, and misguided data exploration (Section 4.3). At the end, I reflect on the design and research implications underlying these concerns and draw a connection between practitioners' concerns and researchers' concerns (Section 4.4).

The contents of this chapter are adapted from a 2020 IEEE VIS short paper I coauthored with Endert and Stasko [79].

4.1 Methodology

Here, I first describe the process of participant recruitment, interviews, and data analysis.

4.1.1 Participants

I interviewed 23 participants (9 female, 14 male) from 19 organizations. Interviewees worked in 12 job sectors including consulting, retail, and education. The organizations ranged from sole entrepreneur, to start-up companies with less than 10 people, to large corporations with more than 100,000 people. The locations of the organizations spanned 5 states in the US (Georgia, Massachusetts, North Carolina, Pennsylvania, and Tennessee). Figure 4.1 provides the detailed demographics of the interviewees.

As an inclusion criterion, I recruited people who employed end-user visualization platforms in their day-to-day work. Interviewees utilized a wide variety of visualization platforms including, among others, Tableau (23/23 interviewees), Power BI (10/23), Qlik (4/23), and Cognos (4/23). They had 2 - 20 years of experience with these systems.

I recruited interviewees through multiple channels. I went to a Power BI user group meeting in Alpharetta and four Tableau user group meetings in Atlanta, Charlotte, and

Philadelphia. During the user group meetings, I approached attendees in order to collect contact information to schedule interviews. I also emailed contacts in my professional networks and potential interviewees found through websites such as LinkedIn. Several interviewees helped identify colleagues and friends in their networks who they thought might be eligible for the study. Interviewees were not compensated.

ID	Gender	Sector	Role	End-User Visualization System(s)	Yrs Exp
I1	F	Accounting	Consultant	Tableau	6
I2	M	Consulting	Executive	Tableau	8
I3	M	Consulting	Consultant	Power BI, Tableau, Qlik	5
I4	F	Education	Librarian	Tableau, Voyant	4
I5	F	Education	Institutional researcher	Tableau	4
I6	M	Education	Communications manager	Tableau	5
I7	M	Education	Data analytics director	Power BI, Tableau, Qlik	7
I8	F	Government	Compliance officer	Power BI, Tableau, Qlik	4
I9	M	Healthcare	Business intelligence developer	Cognos, Tableau	7
I10	F	Healthcare	Revenue cycle analyst	Tableau, VitalStats	3
I11	M	Information technology	Data analyst	Tableau	7
I12	M	Insurance	Engineering manager	Power BI, Tableau	5
I13	M	Internet	Data quality manager	Cognos, SSRS, Tableau	7
I14	M	Internet	Business intelligence developer	Actuate, SAP, Tableau	5
I15	M	Manufacturing	Innovation manager	Tableau	2
I16	F	Manufacturing	Sales analyst	Power BI, Tableau	4
I17	F	Manufacturing	Market analyst	Power BI, Tableau	4
I18	M	Marketing	Senior manager	Cognos, Power BI, Microstrategy, Spotfire, Tableau, Qlik	10
I19	M	Research	Political science researcher	ArcGIS, Tableau, Voyant	>10
I20	F	Research	Research director	Tableau	6
I21	M	Retail	Automation manager	Cognos, Power BI, Tableau	5
I22	F	Retail	Financial analyst	Power BI, Tableau	6
I23	M	Retail	Human resource analyst	Power BI, Tableau, Visier	4

Figure 4.1: Demographics of interviewees. The table shows their job sectors, roles, end-user visualization systems used at work, and years of experience with these systems. Interviewees are sorted by job sector.

4.1.2 Interviews

I conducted semi-structured interviews with the 23 practitioners between November 2019 and February 2020. The interviews were one-on-one except an interview with 2 interviewees. I interviewed the practitioners in their office (1/23), in researchers' lab (2/23), and via video-conferencing software (20/23). The interviews lasted between 40 and 75 minutes. With the permission of interviewees, I recorded the audio for subsequent analysis.

Before the interviews, I prepared a guiding interview script (provided in Appendix A). During the interviews, I ensured that a core set of questions in the script were addressed

while bringing up follow-up questions as interviewees touched on interesting topics. The questions asked concerns three topics: interviewees' data analysis workflow, experiences with data insights, and concerns about using auto-insight systems. For the scope of this chapter, I focus on reporting interviewees' potential concerns about auto-insight systems.

4.1.3 Auto-Insight Prototypes

To probe interviewees' reactions to auto-insight systems, I created three auto-insight systems and demonstrated these systems during the interviews. These systems were modeled on existing commercial products. I counterbalanced their presentation order during the interviews to avoid potential bias due to order effects. In the following, I briefly describe the three auto-insight systems. A demo video of the systems can be found at `https://youtu.be/GKZ4ZMR1ABQ`.

Insight Dashboard. This system (Figure 4.2a) is modeled on the products offered by Narrative Science [103]. It has a dashboard on the right and an automated analysis pane on the left. The automated analysis pane provides textual descriptions of the visualizations to facilitate interpretation. The system updates the textual descriptions as users select a metric or a state.

Insight Page. This system (Figure 4.2b) is modeled on Quick Insights in Power BI [38]. Upon loading the data, the system displays a page of statistical information for user review. Users can add a visualization they consider interesting to an editable dashboard view.

Insight Query. This system (Figure 4.2c) is modeled on Tableau [146]. Akin to Tableau, users can drag and drop dimensions and measures to the encoding shelves to construct visualizations. As users select variable(s), the "Look Ahead" pane displays statistical information (e.g., high correlation and upward temporal trends) that are related to the selected attributes.

Figure 4.2: Auto-insight systems demonstrated during the interviews. Insight Dashboard provides textual descriptions of the visualization in a dashboard (a). Insight Page displays a page of statistical information for review (b). Insight Query enables users to select and drag variable(s), and statistical information is generated in the "Look Ahead" pane (c).

4.1.4 Analysis

I manually transcribed the audios recorded during the interviews. I then segmented the transcripts into passages (most of them are single sentences) and applied open coding. The transcripts were coded throughout the interview study.

During the coding process, I followed constant comparison and theoretical sampling in grounded theory [28]. When coding a passage, I returned to the data repeatedly to compare it with other passages (constant comparison). I labelled related passages as the same category and grouped related categories into a dimension. For each dimension, I considered other possible categories that had not yet emerged in the coded data and focused on these categories in subsequent interviews and coding (theoretical sampling). I iteratively refined the categories and dimensions through frequent discussions with colleagues.

4.2 Who Are These Visualization Users?

Before presenting the concerns mentioned by interviewees, I first describe their data analysis workflow to provide context for the findings.

Reporting was a central activity in the workflow of interviewees. All interviewees employed end-user visualization platforms for crafting reports. Besides traditional reports (e.g., PDFs and presentation slides), most interviewees mentioned creating visualization dashboards (22/23). Audiences of the reports varied. They could be internal audiences in other departments (14/23) and clients (4/23). Interviewees also mentioned creating dashboards for audiences external to the organizations who were not clients (6/23). For example, institutional research departments at colleges created dashboards for prospective students and parents (2/23).

Interviewees played dual roles in the process of report creation: data analyst and visualization designer.

As an analyst, interviewees conducted the analysis required for creating reports (23/23).

Analysis for creating reports often involved measuring key performance indicators of some processes (15/23). Before creating reports, some interviewees conducted exploratory analysis to get familiar with the data (10/23) or look for interesting patterns (10/23). Spreadsheet applications such as Excel were the *"go-to tool"* for such initial analysis (15/23). Interviewees often used Excel for inspecting tables (8/23) and creating pivot tables (7/23). When reporting statistical analysis, interviewees conducted the analysis using statistical software such as R and SPSS (3/23).

As a designer, interviewees designed dashboards for their audience (22/23). They often employed human-centered design [106] for dashboard development (16/23). During the design of a dashboard, they first gathered requirements from end users. Some mentioned receiving a report specification from the end users (7/23). Design requirements might also be a product of discussions between interviewees and their end users (13/23). Throughout the design process, interviewees often demonstrated dashboard designs to the end users for feedback and iteratively refined the design accordingly (15/23).

4.3 Potential Concerns About Using Auto-Insight Systems

In this section, I summarize the five concerns about using auto-insight systems that emerged from the analysis of interview scripts: misinterpretation, non-transparency, loss of agency, information overloading, and misguided data exploration.

4.3.1 Misinterpretation

The most common concern about using auto-insight systems is user misinterpretation (10/23). Several interviewees commented that automation was *"not completely error-free"* (I10) and might produce insights that did *"not make sense"* (I10) (7/23). As interviewees often designed dashboards for some end users (22/23), some were concerned that auto-insights could mislead their users. I12 was a communications manager in a college and created dashboards to showcase research publications. He was worried that auto-insight systems

might not provide a correct interpretation for the end users: *"Yes, I would use it [the auto-insight system], but I'm always wary of is it giving you the correct interpretation of the data that you want the end user wants."*

Some interviewees were concerned that auto-insights were unreliable (8/23). Based on their experience in data analysis, they described several sources from which the unreliability could arise, including poor data quality, violation of assumptions, causal inference, and lack of domain knowledge.

I11, a data quality manager in an internet company, cited poor data quality as a reason why auto-insight systems might generate nonsensical statistical information: *"Might be some limitations just with data cleaning of course. And I think that's always the hardest part of any data set. Just because you have a table and data doesn't mean you can quickly plug it into something because it has many issues: formatting issues, column issues, etc."*

I3, an innovation manager, had been educating the workforce on better analytic practices. He suggested that auto-insight systems often utilized models for generating some conclusions, and these systems might fell apart when the assumptions were not satisfied:

> *I think that [the auto-insight system] is useful. I would want to, from an end-user perspective, make that really relevant because there's a lot of different ways to modeling and forecasting. [...] I think it's something that people often forget about [...] Any time you use an equation, there's an underlying set of assumptions that have to be satisfied for its relevance. Otherwise. you're just going to get a number that's not meaningful or misleading.*

He further emphasized the importance of protecting users from being misled by the auto-insights, explaining, *"You'd want some ways to protect the end users and help end users to ultimately do the right thing without having to be a subject matter expert on it and then overtime they will learn."*

Users might also misinterpret auto-insights when analyzing the causes of an outcome. I3 commented that, in practice, root cause analysis from observational data was challeng-

ing, and people often resorted to correlational analysis when investigating root causes. He said, *"More complex of a business process, the less you can rely on causation. It's more about correlations unless you just want to spend a lot of time and money and other things."* However, correlation is not causation. When users want to understand the causes, but correlation is the evidence they can rely on, users might misinterpret statistical information about correlation as causation. I3 provided an example, saying, *"There's the very commonly used idea of there's a correlation between shark attacks and ice-cream consumption [...] but eating ice cream doesn't cause people to get shark attacks more"* (I3).

Some interviewees also felt that auto-insight systems lacked domain knowledge (4/23). As a result, these systems could provide interpretations that did not make sense. I23 worked at a non-government organization that publicized event data about international armed conflicts. She noted that a correct interpretation of their data required significant expertise beyond the purview of automation:

> *I would be worried that it would be suggesting interpretations that are not necessarily meaningful or not make sense. When it's a bit more subjective like when you enter the world of Social Sciences, that's a trickier line to walk. For example, we code event type, and we have different event type categorizations [...] like armed clashes [...] like violence against civilians [...] like protests [...] And we will code the actors involved like the primary two actors in addition to associate ones [...] But, we don't attribute directionality in all of these events. So, if it's an armed clash, we don't code whose fault it was, who started it, who the perpetrator is [...] We have found information that can be quite biased in reporting. It really depends on what media you look at. If it is a pro-government media, they will always allege that the other side started [the armed clash]. We, as the data creators, know the nuance to apply, but I would be worried if it's just like a machine looking at the data.*

I7 and I12 also stated that auto-insight systems could aggregate the data incorrectly

when they lacked an understanding of the semantics of data attributes, resulting in a misleading conclusion about the data. As an example, it would be unreasonable to sum over the monthly weights of a person to find the total. Not knowing that the values are weights, auto-insight systems might compute the total, and this does not make sense. In 17's words:

> *If you're going to capture where a number is over the course of time where you would basically say a certain measure would not be additive, you couldn't add all the rows to create a logical number. That's where stuff like this tends to fall apart. So, you have to be thoughtful about making sure that you've got the data that is aggregated in logical ways before you start to show this [auto-insight].*

4.3.2 Non-Transparency

Transparency in data observations implies being able to drill down into the observations, ask follow-up questions about them, and validate their correctness. Such transparency is instrumental in building trust with the audience during presentations. Validation also requires transparency and is important because of the financial costs of drawing wrong conclusions from data.

During presentations, non-transparency in data findings can lead to distrust in the audience. I11 commented on the skepticism of the audience when presenting data findings:

> *In many cases, what I've seen is we, as analysts, would come up with certain insights and we would present them to a team of stakeholders. These could be financial folks. These could be statisticians. They could be analysts from other teams [...] In order to build that trust in what we're trying to present, they have questions that are: How did you come to this analysis? What things did you filter out? What things did you include? What did you compare this against? And did you compare this over here as well?*

Interviewees wanted to effectively present data findings and were cautious about reporting findings that were potentially incorrectly (4/23). I16 said, *"I'm definitely conscious*

all the time of accuracy and [...] I'm not going to put anything out there if I don't have this understanding and trust."

Furthermore, incorrect conclusions from data could incur financial consequences. I9, a revenue cycle analyst working in an urgent care center, was responsible for financial reporting. She described that financial data were fraught with errors and data validation with multiple sources was important:

> *Validation is very important to me because we need to see where the money is at. If you are in the EPIC system [the patient financial system], and EPIC says that you created 30 million dollars in charges for January [...] Now, you can go in Tableau, and you can set that up to automatically run at the end of the month [...] and tell you how much charges you made [...] If EPIC says it's 30 million dollars but then my Tableau report says it's 20 million, something's wrong. Where's the extra ten million in my EPIC report? I got to find out what happened to my report. Did I miss something? Did I exclude something from the table to cause it to say 20 million? [...] The business got to pay taxes on stuff like that.*

Such a need for accuracy implies that being able to verify data observations is crucial. Yet, many auto-insight systems are opaque, rendering data validation difficult [29, 98]. This opaqueness was a shared concern among some interviewees (8/23): *"Because I didn't calculate them [the statistical information from auto-insight systems] that means that I'm trusting that the program has calculated them correctly so being able to verify that is kind of important"* (I8), and *"I think one of the problems you might run into is the explanation of how that insight [auto-insight] was garnered. It might be helpful to have a mathematical explanation [...] that says this was filtered out or this was computed by using this"* (I11).

I6, a business intelligence developer, recalled his interactions with the end users and felt that not everyone would appreciate the automation because it could be non-transparent and, therefore, difficult to trust:

The only questionable part I would say, from talking to another person, is that they will want to know what's driving everything. They will want to know if all that's correct. [...] Some people I know of would not trust that [auto-insight]. They would want to verify what the insight is telling them. They would want to find a way to verify if the insight is correct.

Besides the end users, as an analyst, I6 expressed unease in trusting auto-insights himself, explaining, *"Because if I have to present information to a manager, then if the manager asks me a question, I have to be able to explain what the tool is saying and for me in order to do that I would have to understand how the functionality works."*

4.3.3 Loss of Agency

Tukey likened data exploration to detective work [149]: During data exploration, the analyst is motivated by some high-level concerns, engages in opportunistic discovery, and iteratively refines the analytic questions in response to the data observations [3]. Some interviewees similarly preferred a flexible system that enables them to ask follow-up questions (8/23). I13, a political science researcher, described a data exploration scenario where he would ask a series of questions:

I would like to be able to drill in and see the data. If I have an outlier, I want to see the data. I want to see why it is an outlier. [...] Let's say sales and share for Russia are an outlier for this computer. I want to see when it's an outlier. I would want to see the data from that point, but also how far away it is from the mean, a z-score or something like that.

However, many auto-insight systems lack flexibility. For instance, some systems (e.g., DataShot [156] and Calliope [131]) communicate insights by automatically generating a fact sheet of statistical information. These systems essentially turn data exploration into a review process that precludes users from drilling down into an auto-insight. Several

interviewees commented that a limited ability to dig deeper reduced the utility of auto-insight systems. I8 said, *"One of the biggest limitations is that this [auto-insight system] answers certain questions but there's bound to be another question."* I15 shared a similar concern:

> *I think it helps the workflow in terms of data exploration standpoint and [...] see where there might be some outliers, averages, correlations, etc. Then, the limitations come into making it useful for the end user and knowing what the end user is looking for, like what questions are you actually trying to answer. I don't think it necessarily will be who has the highest average. I think there will be some more sophisticated questions. It's where it [auto-insight system] will fall short.*

Besides limiting the freedom of drilling down, auto-insight systems could reduce user agency by indulging reliance on the recommendations: With the recommended statistical information, users may be more inclined to blindly follow these recommendations, as opposed to taking the initiative to dive into different aspects of the data. I16 hinted at the potential for auto-insight systems to discourage data exploration:

> *I do think that while those types of tools [auto-insight systems] are very cool for end user folks, sometimes, I would want analysts to be encouraged at all times to play and to explore [the data]. I think [data exploration] is something to encourage and those types of things [auto-insight system] somehow discouraged that.*

I1 even felt that auto-insight systems could be manipulative during data exploration:

> *For me, if a tool just does everything for me, I don't know how I feel about it because I would feel like I'm so guided by the tool I'm almost manipulated by the tool. I'm like why you don't let me think how I want to look into it. How do I know you're giving me the best result? [...] I think it's very important to*

put humans as a center of research or exploration [...] Giving the researchers the power and the flexibility to go down the direction they want to go to and to explore the aspect they want to explore. I think it is really important.

4.3.4 Information Overloading

Another concern surrounds information overloading. Many auto-insight systems enumerate numerous patterns in the data and present a large amount of potentially interesting statistical information for user review. Several interviewees discussed how auto-insight systems could overwhelm users:

I think right now it could be a scenario of analysis paralysis where it [auto-insight system] can be overwhelming for some. (I4)

I think the limitation was the information overload part where it [auto-insight system] is looking at everything in it. Sometimes, I use this analogy with my clients where you go to the cheesecake factory and the menu is like 50 pages. You just can't decide what you want because it's just way too much. I feel like for your clients, one of your jobs as their analysts is to give them the two or three things that they should focus on that are most important. I almost think the automated insight needs to be further refined by their analyst before it goes to the end users. (I5)

There's kind of like a fine line or a point of no return that you cross as a data vis developer where you just put so much content out there that no one really knows what to look at anymore. (I14)

4.3.5 Misguided Data Exploration

Finally, I8 found the potential for auto-insight systems to misguide users during data exploration worrisome: *"If I see data already saying that this is the number to look at, is*

that gonna guide my analyses? Am I gonna go down a rabbit hole that maybe I shouldn't because this program has told me to? Am I going to get distracted by that number when really the question was in the other direction all along?"

4.4 Discussion

This chapter has presented five potential concerns about the use of auto-insight systems: misinterpretation, non-transparency, loss of agency, information overloading, and misguided data exploration. Now that we know visualization users' concerns. How do we make auto-insight systems more usable to them? In this section, I comment on several research and design implications for the findings, connect practitioners' concerns to researchers' concerns, and reflect on the limitations of the interview study.

4.4.1 Investigating the Misleading Power of Auto-Insights

A potential concern to many interviewees was misinterpretation. In the field of data visualization, researchers have been investigating how visualizations shape data interpretation. They found that a wide range of factors could influence the messages communicated to viewers. In particular, users often apply domain knowledge when viewing a visualization, and this knowledge could prime a viewer to obtain a particular message from a visualization [167]. Moreover, perceptual and cognitive biases play a role in manipulating data interpretation [37]. For example, distorting the aspect ratio of a line chart can lead to an inaccurate assessment of trends in the data [54]; when viewing a scatterplots, the accuracy of class separability judgment depends on the scatterplots users saw before [151].

In spite of the research to understand the impact of visualization design on data interpretation, we know relatively less about the impact of auto-insights. Auto-insight systems could be characterized by their proactive behaviors. While in typical data analysis, users actively glean insights from visualizations, users are passively told about noteworthy statistical information when using auto-insight systems. Such proactive behaviors imply that

auto-insight systems could exert undue influence on user interpretation. The potential undesirable influence on users and their recent deployment in widely-adopted commercial systems [145, 38] call for a better understanding of whether auto-insights could mislead users, and if they could, how to protect users from being misled. Using auto-insights about causation as a case study, the next chapter hopes to shed light on answers to these questions.

4.4.2 Improving Transparency in Auto-Insight Systems

I have learnt that transparency in auto-insight systems can inspire trust in the audience and facilitate data validation. It is therefore desirable to reduce non-transparency in auto-insight systems, and one approach is to provide explanations. In designing explanations for auto-insight systems, lessons about what explanations to provide could be learnt from other research areas. For example, Lim et al. [89] proposed different types of explanations for context-aware systems (e.g., what, why, why not, what if, and how to). In collaborative filtering, Herlocker et al. [57] derived explanations based on the processes underlying a collaborative filtering system.

Similarly, auto-insight systems could provide textual statements that explain the process of generating the recommendations and the limitations. For instance, an auto-insight system may recommend a scatterplot that shows a high correlation between sales and discount. The system could describe the generation step: Missing values are imputed, and the chart is considered to be potentially interesting because the Pearson correlation is higher than 0.4. To explain the limitations, the system could further highlight the caveat that there are outliers in the data, and Pearson correlation is not robust in the presence of outliers.

To support data validation, validation mechanisms could be incorporated into auto-insight systems (e.g., [82, 172]). Zhao et al. [172] envisioned a visualization system that can automatically infer the hypotheses users have tested during visual data analysis and conduct statistical tests on the background to protect users from false discovery. Another design idea is to provide ways for testing model and metric assumptions. For instance,

Pearson correlation relies on the absence of outliers. Auto-insight systems utilizing Pearson correlation could offer methods for identifying outliers to ensure that the assumption is met.

I note, however, that supporting a higher level of transparency is not an unmitigated good. First, providing more information through explanations increases cognitive load when users are already deluged with much information during data analysis. Prior research indicated that providing too much transparency in an algorithmic grading system can cause students to trust the algorithm less [71]. Furthermore, explanations could expose the weaknesses of an automated system, leading users to come to a realization that the algorithm is not as powerful as they thought [44].

The extent to which these research findings apply to auto-insight systems warrants further investigation. Nevertheless, it seems reasonable to conclude that transparency support for auto-insight systems should be carefully designed so that users are willing to trust and utilize these systems for real-world tasks. What is the right amount of transparency? What information should an auto-insight system provide to enhance transparency? What is the appropriate presentation of explanations? Addressing these questions is ripe for further research.

4.4.3 Promoting Relevance to Avoid Information Overloading

North identified relevance as a characteristic of data insight [107]. In his words, "Insight is deeply embedded in the data domain, connecting the data to existing domain knowledge and giving it relevant meaning. It goes beyond dry data analysis, to relevant domain impact" [107]. One way to avoid information overloading is to promote relevance in the auto-insights. I7 suggested that the auto-insights could be ranked by relevance to users: *"You want to make sure that you are showing kind of the most important to the least important."* Yet, he also noted the challenges in getting the ranking right because ranking requires inference of user interests: *"You're going to have to know a little bit about what the person is trying to understand. So, I think the order in which you present the narrative*

can be challenging to get right."

As opposed to inferring user interests, an alternative idea for promoting relevance in auto-insights is to directly acquire the mental models (e.g., expectations and hypotheses about the data) from users. For instance, Choi et al. [16] proposed concept-driven visual analytics. They envisioned a system that enables users to externalize their expectations about the data through natural language (e.g., I expect the US to have the highest GDP per capita in the world). The system will then recommend relevant visualizations to users (e.g., showing a bar chart that reveals the GDP per capita of different countries with a textual description that indicates the true rank of the US). Richer information about a user's mental model could increase the likelihood that users consider the recommended visualizations and data facts to be relevant.

4.4.4 Balancing Agency and Automation

Interviewees were concerned about the loss of agency when using auto-insight systems. In typical visual data analysis, users manually create visualizations and interpret the visualizations to glean insights. With auto-insight systems, users delegate much control to the automation as the system automatically creates visualizations and extracts statistical information from data. Is this the right balance between agency and automation? Heer [53] suggests that an appropriate balance should be asymmetric: "Automated methods suggest possible actions, which are then displayed for review and revision by the user, who remains the ultimate decision maker." This describes the way many auto-insight systems (e.g., [32, 60]) work—they recommend visualizations and data facts and allow users to decide whether to attend to the recommendations. Can we therefore conclude that these systems have achieved the right balance between agency and automation?

From interviewees' feedback, however, providing recommendations and letting users choose whether to pay attention to them do not seem to guarantee the right balance. Some interviewees were concerned that recommendations from auto-insight systems could in-

dulge reliance and discourage data exploration. Reliance on recommendations seems to imply that the automation encroaches upon user agency. However, whether this reliance constitutes an undue influence to users is disputable. The research community will benefit from further user studies and extended debates to learn about the fundamental nature of recommendations: Do recommendations inherently encourage reliance? Is reliance a problem? Do recommendations inevitably present an intrusion into user agency?

Besides learning about the fundamental nature of recommendations, studying different recommendation approaches could inform the design for a right balance between agency and automation. One consideration in designing recommender systems is whether to offer recommendations reactively (provide suggestions specific to an initiating interaction) or proactively (provide suggestions even when users do not request them).

Reactive and proactive recommendations appear to produce different perceptions and induce different behaviors [52, 165]. However, results from user studies are sometimes contradictory. For example, in studying text editor assistant, Jun et al. [165] found that users had a low expectation on the quality of proactive recommendations and often considered them useful. In contrast, Guo et al. [52] studied proactive recommendations in a data wrangling system and found that users often wanted to maintain initiative and found the proactive behavior distracting. The contradictory results could imply that user perceptions depend on personality (e.g., willingness of users to follow recommendations), recommendation quality, and the context of use. Such a diversity of factors make an appropriate balance between agency and automation hard to design for.

The final concern about the use of auto-insight systems is misguided data exploration. While some visual analytics researchers have advocated for offering automated guidance to users during data analysis [20], we miss a discussion of misguidance—guidance can be excessive, harmful, biased, and purely wrong. The guidance model offered by Ceneda et al. [20] is a starting point for studying the intricacies of misguidance. Their model comprises different types of guidance (data, tasks, visual analytics methods, users, and infras-

tructure) and different degrees of guidance (prescribing, directing, and orienting). Studying the impact of misguidance on data exploration could be grounded in these dimensions. For example, when a system provides *prescribing* guidance on *visual analytic methods*, how does the quality of guidance affect the perceptions of the system and the quality of data exploration? Investigations like this could help resolve the quandary about when and how automated guidance turns into misguidance.

4.4.5 Connecting Practitioners' Concerns and Researchers' Concerns

Some interviewees were concerned about misinterpretation. This concern is shared by both Correll [29] and McNutt et al. [98]. Both are worried that auto-insight systems could act as a "p-hacking machine" that enumerates numerous patterns in data and identifies noteworthy ones without validation. In essence, the "p-hacking" problem lies in sampling variability: Auto-insight systems could be unreliable because drawing conclusions about a population from a data sample involves uncertainty. Aside from sampling variability, this interview study revealed other sources of unreliability in auto-insight systems, namely poor data quality, violation of assumptions, causal inference, and lack of domain knowledge.

Correll [29] and McNutt et al. [98] have similarly considered non-transparency and loss of agency as potential issues in auto-insight systems. Concerning information overloading, the visualization community has recognized it as a challenge [91]. However, information overloading is less emphasized in the literature when it comes to auto-insight systems. Finally, despite the ongoing effort to investigate automated guidance in visual analytics [20], we lack a discussion about the impact of misguidance (i.e., excessive, harmful, biased, and even purely wrong guidance) during data analysis. My interview results could enrich the discourse on the potential pitfalls of auto-insight systems by providing empirical basis for confirming the researchers' concerns and identifying new perspectives for considering such pitfalls.

4.4.6 Study Limitations

This study suffered from the same limitations as typical interview studies. First, interviewees were asked to recall past experiences. Their depictions might be imprecise due to a limited ability to recall the past [108]. Moreover, the number of interviewees who mentioned a view or an activity does not reflect the frequency of the view or activity among the whole population of visualization users. Surveys are better suited for quantifying a phenomenon [108]. I also recognize that the potential concerns I identified may not be complete. Additional interviews with a broader set of visualization users could augment the findings. Finally, Tableau users might be overrepresented in our study. In future studies, researchers could recruit a more representative set of interviewees based on surveys of visual analytics systems (e.g., [170]) and their market shares.

CHAPTER 5

PERCEPTION OF CAUSALITY IN AUTOMATED DATA INSIGHTS

In the previous chapter, I discussed that misinterpretation of auto-insights constitutes a potential concern to visualization users. However, in reality, do auto-insights mislead users? This chapter presents an investigation of the misleading power of auto-insights using auto-insights about causation as a case study.

Why causation? My research was motivated by the deployment of Explain Data in Tableau [145]. Explain Data is an auto-insight functionality that provides suggestions about the potential cause of an extreme value in data. For example, when exploring a data set about the states in the US, a user may observe that students in Massachusetts have the highest average ACT Math score. With a simple click, she can ask Explain Data to suggest explanations for the high score.

However, causal inference from observational data (as opposed to randomized experiments) is inherently challenging [123]. In the above example, Explain Data infers that the rate of teenage pregnancy is negatively correlated with ACT Math score and that the low rate of teenage pregnancy in Massachusetts may lead to the high ACT Math score (Figure 5.1). The veracity of this claim is questionable.

Compounding the unreliability is the huge user base of Tableau. Thousands of users could already utilize Explain Data, but we do not understand enough about the misleading power of auto-insights that describe causal claims. I argue that it is important to study two research questions: ***Do people misinterpret auto-insights that describe causal claims? If they do, does providing a simple warning message help ensure a correct interpretation?***

Thus, I conducted two crowdsourced studies to investigate the impact of different information (scatterplot, textual description about correlation, and warning message) shown alongside a causal claim on user perceptions. In both studies, participants reviewed a series

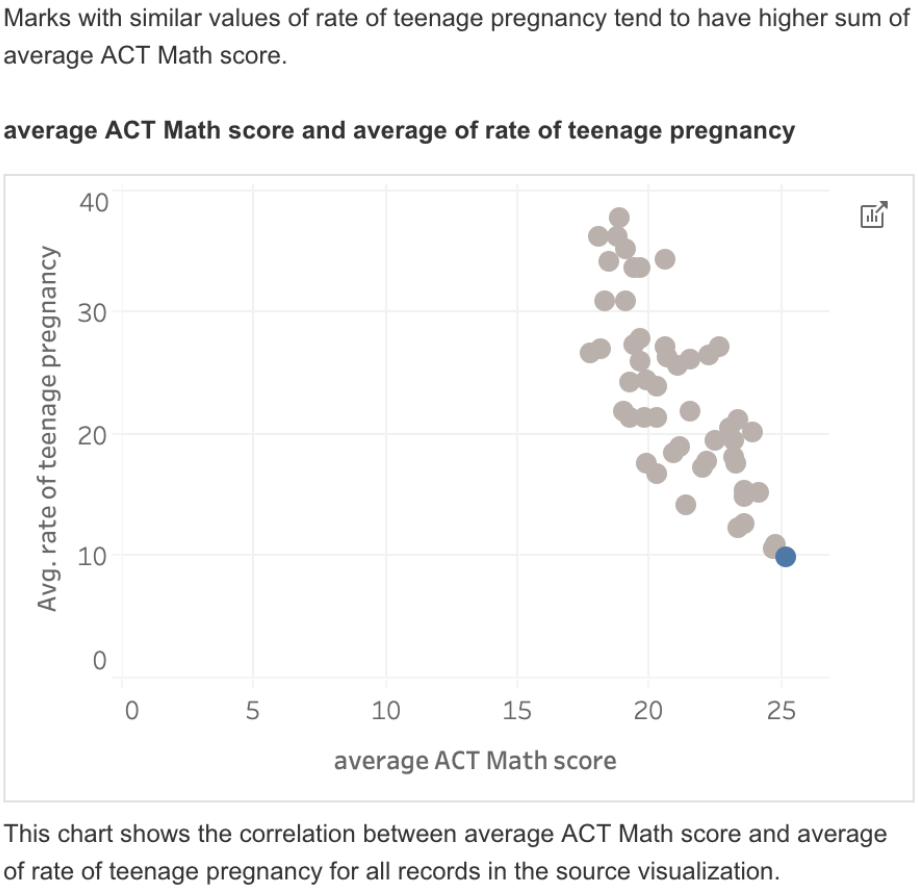


Figure 5.1: An answer generated by Tableau Explain Data [145]. The user asks about the high ACT Math score in Massachusetts. Explain Data infers that teenage pregnancy rate and ACT Math score are negatively correlated, and the low teenage pregnancy rate in Massachusetts might cause the high ACT Math score. It shows the data using a scatterplot in which each dot is a state and the blue dot is Massachusetts.

of answers to why questions (e.g., why students in Massachusetts have high ACT Math scores on average). This setup aimed to emulate the question-answering nature of Explain Data. The answers participants reviewed had different designs. Across the designs, I compared the perceived plausibility of the causal claims, user trust in the system, the awareness of the system's flaws, and users' tendency to associate correlation with causation. While the first study presented answers with different levels of plausibility, the second presented only reasonable answers.

Do people misinterpret auto-insights that describe causal claims? From the first study,

I found that participants tended to disagree less with an unreasonable causal claim when a scatterplot was presented alongside the claim. Also, they agreed more with a causal claim of which the plausibility was difficult to judge with a scatterplot next to the claim. However, I did not find evidence that a textual description about correlation affected the plausibility of a causal claim when it was shown alongside the claim. These results imply that auto-insight systems could utilize the persuasive power of visualizations to create an illusion of causality: Although scatterplots only provide evidence about correlation, they can increase users' tendency to agree with causal claims.

Does providing a simple warning message help ensure a correct interpretation? The first study showed that participants accepted a reasonable causal claim more cautiously when they were shown a simple warning message about the system's potential confusion of correlation and causation. I further observed a general unawareness of "correlation does not imply causation," but the warning seemed to reduce participants' tendency to associate correlation with causation. However, these effects of warning seemed to be unreliable. In the second study where the system only presented reasonable answers (as opposed to a mix of reasonable and unreasonable answers in the first study), I did not observe the effects of warning on the plausibility causal claims and the tendency to associate correlation with causation. Furthermore, even when participants were warned about "correlation is not causation," they tended to agree that a scatterplot that visualizes correlation in the data could imply causation. The unreliability of warning indicates that avoiding misinterpretation would require solutions beyond simple warning messages.

In the remainder of this chapter, I first discuss what it means by "correlation is not causation" and people's vulnerability to the illusion of causality (Section 5.1). Through a literature review, I then argue that there is gap in our knowledge about causal illusion when using auto-insight systems like Explain Data (Section 5.2). Next, I describe a series of crowdsourced studies that investigated the misleading power of auto-insights about causation and the effectiveness of warning to inoculate users from being misled (Section 5.3

– Section 5.5). Finally, I discuss the importance of skepticism when utilizing auto-insight systems that generate causal claims and propose design ideas to encourage skepticism (Section 5.6).

The contents of this chapter are adapted from a 2021 SIGCHI full paper I co-authored with Lo, Endert, Stasko, and Qu [80].

5.1 Correlation Is Not Causation

My objective is to investigate whether auto-insights that describe causation could mislead users. One way auto-insight systems could mislead users is to provide correlational evidence to induce an illusion of causal relationship. Such correlational evidence could be misleading because correlation is not causation. As a reader, you may be wondering: What is causation? What does "correlation is not causation" mean? How does correlation create causal illusion? In the following, I consider each question in turn.

5.1.1 What Is Causation?

Causation can be defined in two senses: theoretically and experimentally. The theoretical definition of causation is based on the notion of counterfactual outcomes [58]. Consider a patient who received a heart transplant and passed away. If, by some divine revelation, we know that the patient would not have passed away if he had not received the transplant, we say that the transplant has a causal effect for the patient. This reasoning involves comparing the outcome when an action is taken versus the counterfactual outcome if the action had withheld. It is theoretical because of the impossibility to observe both outcomes on the same patient.

The experimental definition of causation is based on the notion of ceteris paribus (all other things being equal) [139]. In a simple experimental setup, scientists have two identical entities (e.g., identical twins) or two statistically identical samples. They then randomly assign one as the treatment group and another as the control group. Because all things other

than the treatment are the same across the two groups, any differences in the outcome can be attributable to the treatment. Whereas the theoretical definition considers applying both treatment and control to an individual, the experimental definition considers applying either treatment or control to an individual.

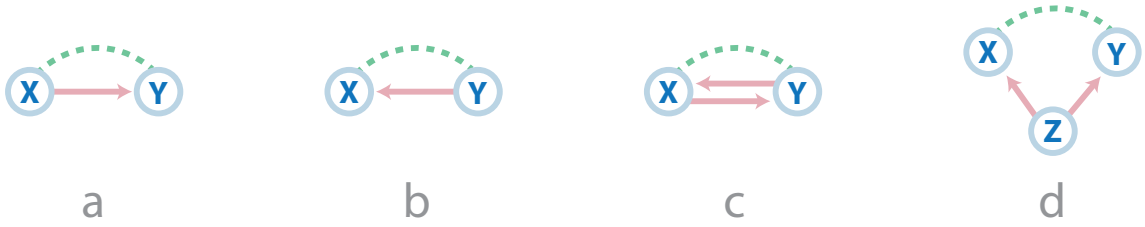


Figure 5.2: Four reasons for correlation between two variables X and Y. In the causal graphs, a directed edge indicates causation, and a dotted edge indicates correlation. Variable X is a direct cause of a variable Y (a). Variable Y is a direct cause of variable X (b). Mutual causality between variables X and Y (c). Variables X and Y are not causally related, but a third variable Z is a direct cause of both variables X and Y, thereby creating a correlation between them (d).

5.1.2 What Does "Correlation Is Not Causation" Mean?

Correlation does not imply causation because correlation can be produced by reasons other than causation. Statisticians have proposed four reasons of correlations between two variables [139]. These reasons can be summarized by the causal graphs in Figure 5.2. In the causal graphs, a vertex is a variable, a directed edge represents causation, and a dotted edge represents correlation. The first reason for correlation is that a variable X is a direct cause of a variable Y (Figure 5.2a). Similar, variable Y can be a direct cause of variable X (Figure 5.2b). Mutual causality can also occur between the variables (Figure 5.2c). For example, an increase in police spending may lead to a decrease in crime rate. Meanwhile, a decrease in crime rate may subsequently lead to a decrease in police spending.

Besides the three causal relationships, correlation can be a result of a third variable Z (Figure 5.2d). Although variables X and Y are not causally related, variable Z, being a direct cause of both, causes variables X and Y to move together. Statisticians often refer to the presence of a third variable that confounds the causal relationship between two other

variables as omitted variable bias [139]. For example, life expectancy and the number of TVs at home likely do not have a causal relationship. Nevertheless, they might correlate because of wealth: As people get richer, they live longer and have more TVs at home. Despite this correlation, it would be ridiculous to conclude that more TVs at home leads to a longer life. A common third variable is time. Time might lead to natural fluctuation in some variables that are not causally related. For example, towards the end of a year, the temperature decreases and the monthly sales of Christmas trees increases. However, an increase in Christmas tree sales does not lead to a decrease in temperature.

5.1.3 How Does Correlation Create Causal Illusion?

While correlation does not imply causation, it is easy to confuse them, leading to illusions of causality [97]. Illusions of causality occur "when people develop the belief that there is causal connection between two events that are actually unrelated" [97]. Such illusions are common in daily life and often cultivated superstitious thinking [97]. For example, bad things might happen after you break a mirror, and you might believe that breaking a mirror causes bad luck [33]. Scientists have found that various factors including the probability of the cause, the probability of the outcome, whether a person is depressed, and whether there are alternative causes can influence the causal perception between two events [97]. In this chapter, I investigate whether data analysis systems can induce an illusion of causality by providing correlational evidence.

5.2 Could Auto-Insights about Causation Mislead Users?

I hypothesize that auto-insight systems could provide correlational evidence to create an illusion of causality. In data visualization, researchers have investigated a wide range of factors such as knowledge, perceptual biases, and cognitive biases that can shape user interpretation of data. Such findings provide indirect evidence of the vulnerability of users to data misinterpretation. Here, I summarize some of these findings and highlight the research

gap related to my dissertation.

Knowledge external to visualizations often affects how we interpret visualizations. As users look at a visualization, they often apply their domain knowledge [125]. Xiong et al. [167] showed that this prior knowledge could prime a viewer to obtain a particular message from a visualization and lead the viewer to believe that other viewers would receive the same message. Besides prior knowledge, social information also affects data interpretation. Kim et al. [70] found that seeing others' expectations about the data influenced people's trust in the accuracy of the data.

Moreover, perceptual biases play a role in manipulating data interpretation [31, 111]. For example, distorting the aspect ratio of a line chart can lead to an inaccurate assessment of trends in the data [54]; truncating the y-axis in a bar chart exaggerates effect sizes [30]; the neighborhood of a bar in a bar chart can change the perceptions of the bar's height [171]. However, these biases could be mitigated through judicious design. For instance, Ritchie et al. [122] showed that an animated transition from an untruncated bar chart to a truncated one could avoid misinterpretation.

Cognitive biases can further change the lens through which we interpret visualizations [37]. An example is priming and anchoring effects. Calero Valdez et al. [151] conducted experiments to show that the judgment of class separability in scatterplots depended on the scatterplots users saw before. Biases in data interpretation can have further consequences on decision making. Dimara et al. [36] provided evidence that the presence of dominated data points in a scatterplot influenced the judgement of which data points were dominating.

Besides knowledge and biases, subtle design choices also matter to data interpretation: Titles can have a misleading impact on visualization interpretation [11, 62, 74, 75]; visual embellishments can affect the insights we gain from visualizations [12, 101].

My work closely relates to Xiong et al.'s [166] experiments that studied how visualization design might lead to causal conclusions from correlation. They concluded that the

level of data aggregation could influence the perceived causality in data.

While these studies provided evidence that information visualizations hold significant power to shape data interpretation, we lack direct evidence concerning the impact of providing correlational evidence on causal perception when using auto-insight systems. This research contributes to the above literature by studying whether correlational evidence shapes causal perception. Specifically, I studied whether two forms of correlational evidence (scatterplot and textual description about correlation) created an illusion of causality and whether simply warning users that correlation is not causation reduced this illusion.

5.3 Pre-Study: Collecting Causal Statements

As a starting point to understand the effects of correlation evidence and warning in the answers to why questions, I focus on why questions about extremum (i.e., an extreme value). An example is why students in Massachusetts have high ACT Math scores on average (Figure 5.1). Finding extremum is a common task during data analysis [4]. Furthermore, functionality to answer such questions has already emerged in commercial systems such as Tableau [145]. Results from this study could offer design guidelines in practice.

In study 1, I showed participants a series of answers to why questions. I created answers with different visual designs and assessed user perceptions of the systems given the designs. Due to the inherent challenges in causal inference [123], auto-insight systems often provide unreasonable answers to why questions. To emulate the behaviors of these systems, I selected causal claims with different levels of plausibility as answers presented to participants. To select these causal claims, I conducted a pre-study.

5.3.1 Methods

Datasets

I planned to generate causal claims that were backed up by observational data and considered using synthetic data. However, my goal was to study user perceptions of auto-insight

systems, and the credibility of the data might affect user perceptions. To control for the potential experimental confounds, I used real-world data instead.

I first curated a dataset about states in the US from sources including the US Census Bureau [150], National Center for Education Statistics [104], and Kaiser Family Foundation [65]. The curated dataset has 258 attributes about demographics, healthcare, and education for each US state. I chose these topics because they are accessible to laypeople. This enabled participants to judge the plausibility of the generated causal claims based on common sense.

With the curated data table, I computed the Pearson correlation for all attribute pairs. To find attribute pairs with a potential causal relationship, I collected the ones with a high correlation (above 0.7 or below -0.7). For each of the 1522 attribute pairs with a high correlation (e.g., employment rate and poverty rate), I found a state (e.g., Mississippi) that has an extreme value for both attributes and omitted the attribute pairs where such a state did not exist.

Based on the collected attribute pairs and states, I generated causal claims (e.g., low employment rate in Mississippi may be a factor that leads to the high poverty rate in Mississippi). The plausibility of this claim can be affected by the plausibility of the causal relationship (e.g., employment rate affects poverty rate) and that of the information about the state (e.g., Mississippi has a high poverty rate). Since I intended to assess the plausibility of the causal relationship, I removed the states from the causal claims (see Figure 5.3).

I carefully picked 30 reasonable claims, 30 unreasonable claims and 30 claims that were hard to tell if they were reasonable (hereafter, *hard-to-tell claims*). I verified and ranked the plausibility of these claims through a study on Amazon Mechanical Turk (MTurk).

Participants

I randomly segmented the 90 claims into five batches of 18 claims and recruited 20 workers on MTurk to rate each batch (100 unique workers in total). I limited the tasks to workers

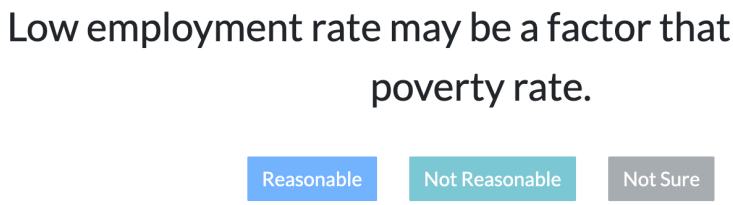


Figure 5.3: Experiment interface used in the pre-study.

in the United States and had an acceptance rate of 95% or above. During data analysis, I omitted participants who failed to pass attention checks (but compensated them for participation). I recruited participants until reaching the target sample size for each batch. Participants were compensated $1 for the study that took approximately 5 – 10 minutes.

Among the 100 participants, 55 were male, and 45 were female. They aged 22 – 64 (M=35.5, SD=11.2). Participants reported their educational attainment to be high school (8 participants), professional school (18), college (49), graduate school (17), PhD (7), and postdoctoral (1).

Procedure

Each participant was randomly assigned to rate one of the five batches of 18 claims. Participants first filled out a demographic survey on their gender, age, and highest education level. They then saw a series of 18 causal claims that were presented on separate pages (Figure 5.3). I randomized the presentation order of these claims to prevent order effects. Based on the plausibility of each claim, participants selected one of the three options: *Reasonable*, *Unreasonable*, and *Not Sure*. As each participant rated more than a dozen causal claims, I used the three options rather than a Likert scale with five options or more to keep the study short. During the study, participants also answered two attention check questions asking them to directly select one of the three options.

5.3.2 Results

For each causal claim, I computed the probabilities that participants selected *Reasonable*, *Unreasonable*, and *Not Sure*. I then calculated the entropy for each claim. A low entropy implies that participants mostly voted for the same option, whereas a high entropy means that participants' votes tended to distribute across the three options. Within each bucket of the 30 reasonable claims, 30 unreasonable claims, and 30 hard-to-tell claims, I ranked the claims by entropy.

For the 30 reasonable claims, I ranked them in increasing order of entropy. The top claims had a low entropy because participants mostly voted for *Reasonable*. For the 30 unreasonable claims, I again ranked them in increasing order of entropy. Participants mostly selected *Unreasonable* for the top claims. For the 30 hard-to-tell claims, I expected that the claims where the plausibility was the most difficult to judge had a high entropy score. This is because participants likely struggled to choose among the options. I sorted these claims in decreasing order of entropy.

Figure 5.4 shows the top three reasonable, unreasonable, and hard-to-tell claims based on the above ranking. I provide a ranked list of all the 90 claims in Appendix B. Study 1 confirmed the validity of the results: In study 1, when participants rated the claims on a 7-point Likert scale, they tended to agree with the top reasonable claims, disagree with the top unreasonable claims, and be neutral about the top hard-to-tell claims.

5.4 Study 1: Providing Answers with Different Plausibility

Having collected the ranked lists of reasonable, unreasonable, and hard-to-tell causal claims, I designed a between-subject experiment during which participants conducted a series of tasks. In each task, participants saw a why question (e.g., why Mississippi has a high poverty rate) and the system's answer (e.g., low employment rate in Mississippi may be a factor that leads to the high poverty rate). The answers were presented differently across

TOP TREEE REASONABLE CLAIMS

R: 20 | U: 0 | N: 0	low employment rate ⟶ high poverty rate
R: 20 | U: 0 | N: 0	high access to exercise opportunities ⟶ high percentage of adults with excellent health
R: 19 | U: 1 | N: 0	high percentage of teens who abuse alcohol ⟶ high percentage of driving deaths with alcohol involvement

TOP THREE UNREASONABLE CLAIMS

R: 1 | U: 19 | N: 0	low percentage of adults who are obese ⟶ high median housing price
R: 2 | U: 18 | N: 0	high percentage of adults age 65+ with all teeth extracted ⟶ low percentage of households with internet access
R: 2 | U: 18 | N: 0	low motor vehicle crash deaths ⟶ high average ACT Math score

TOP THREE HARD-TO-TELL CLAIMS

R: 8 | U: 5 | N: 7	high percentage of adults who smoke ⟶ high percentage of people with vision difficulty
R: 8 | U: 7 | N: 5	high employment rate ⟶ low percentage of adults with poor physical health
R: 7 | U: 5 | N: 8	low percentage of infants who were breastfed ⟶ low life expectancy

Figure 5.4: The top three reasonable, unreasonable, and hard-to-tell causal claims. Each row shows a causal claim (right) and the votes (left). For example, the most reasonable claim was "a low employment rate may be a factor that leads to a high poverty rate." It got 20/20 votes for *Reasonable* (R), 0/20 vote for *Unreasonable* (U), and 0/20 vote for *Not Sure* (N).

conditions. I provide screenshots of the experiment interface in Appendix C.

When designing the presentation of answers, I considered its complexity to typical end users of auto-insight systems. One way to answer why questions (e.g., why Mississippi has a high poverty rate) is causal graph [123, 155], a technique often employed in statistics literature for visualizing complex causal relationships. However, causal graphs may require more advanced statistics training to understand.

I also considered showing multiple factors in the answer but was concerned about introducing experimental confounds. For example, the number, the perceived plausibility, and the underlying causal relationship of the factors could potentially alter the perception of system performance. Yet, using real-world data implied that these variables could be difficult to control for. I therefore adopted a simplified design where the system responded to a why question by stating a factor that could answer the question.

Throughout the study, I described the system under investigation as a "question-answering system" rather than an "auto-insight system." This was because unlike "auto-insight," the

term “question-answering” is more intuitive and could be understood without exposure to visualization research.

5.4.1 Methods

Conditions

Building on prevailing system designs and the research literature, I focused on two types of correlational evidence (scatterplot and textual description about correlation) to investigate whether they created an illusion of causality. I further studied the effectiveness of warning in reducing the illusion. Here, I describe these three types of information:

Scatterplot. Scatterplots are common for showing the relationship between two numerical variables [128]. They have also been applied in Explain Data for showing the relationship between cause and effect (Figure 5.1).

Textual description about correlation. While the causal claim (e.g., low employment rate in Mississippi may be a factor that leads to the high poverty rate) describes a single state in the US, a description about correlation (e.g., as employment rate decreases, poverty rate tends to increase) depicts the overall trends for all the US states. Auto-insight systems often provide such descriptions next to a visualization to facilitate interpretation [136].

Warning message. Prior studies found that although scatterplots and the textual descriptions only reveal correlation, they might induce an illusion of causality [166]. A mitigation strategy is to use a message to warn users that correlation is not causation. While such warning is less common in visualization systems, it is commonly used in other systems (e.g., web browser) to prompt safety-related behaviors (e.g., not to click on phishing websites) [41]. It would be interesting to learn about if a simple statement is enough to raise awareness of the system’s potential flaws and reduce users’ tendency to confuse correlation and causation.

Based on these three information types, I designed four answer interfaces by adding the information types one by one. Participants were randomly assigned to a condition where

the answers adopted one of the four designs:

Claim only (Figure 5.5b). The system only shows a claim about cause and outcome as an answer to a why question.

Claim + vis (Figure 5.5c). Beside the causal claim, the system visualizes the cause and outcome variables using a scatterplot. The aspect ratio of point clouds in a scatterplot can affect correlation estimation [27, 100]. To support a consistent correlation estimation, I controlled the aspect ratio. For each axis, I set the lowest value to be (min value of the data $- 0.15 \times$ range of the data) and the highest value to be (max $+ 0.15 \times$ range).

Claim + vis + description (Figure 5.5d). The system additionally states the correlation between the cause and outcome variables with a textual description.

Claim + vis + description + warning (Figure 5.5e). To encourage users to carefully evaluate a causal claim, the system warns that the scatterplot only shows correlation, and that correlation is not causation.

Participants

A power analysis indicated that for a significance level of 0.05 and a power of 0.8, detecting a medium effect size of $f = 0.25$ using one-way ANOVA required 180 participants (45 participants per condition). As I planned to conduct non-parametric tests (see the Quantitative Measures section), I targeted a slightly larger sample size (200 participants in total or 50 participants per condition) following guidelines on sample size determination for non-parametric tests [84].

During participant recruitment, I limited the study to workers in the United States, had an acceptance rate of 95% or above, and did not participate in the pre-study. The study took approximately 10-20 minutes, and I compensated participants $2.90. At the end, I recruited 200 unique workers on MTurk.

The survey had two interpretation checks for assessing scatterplot comprehension and three open-ended questions (details in the Procedure section). I excluded participants who

Figure 5.5: The four experimental conditions. A user asks about the high poverty rate in Mississippi (a). The system answers only a causal claim (b), shows a scatterplot next to the claim (c), adds a description about the correlation (d), and warns about the system's flaws besides showing the previous information (e).

did not pass any of the interpretation checks or provided gibberish answers for any of the open-ended questions (but compensated them for participation). Overall, the data quality was poor. For example, many participants provided canned responses for some open-ended questions. I omitted 123 participants and continued recruiting until reaching the target sample size.

Participants aged 20 – 69 (M=35.4, SD=10.1). 131 were male, 68 were female, and 1 preferred not to say. They reported different educational attainments: high school (29 participants), professional school (22), college (109), graduate school (35), PhD (1), and post-doctoral (4). Concerning data analysis expertise, 44 had none, 64 were beginners, 71 were intermediate, and 21 were advanced. For experience with visualization platforms (e.g.,

Tableau), 82 had none, 60 were beginners, 36 were intermediate, and 22 were advanced. When asked about the frequency of using question-answering systems, 132 reported never, 30 reported rarely, 23 reported weekly, and 15 reported daily.

Procedure

I first randomly assigned participants to one of the four conditions. For all conditions, the study consisted of five main stages (Figure 5.6).

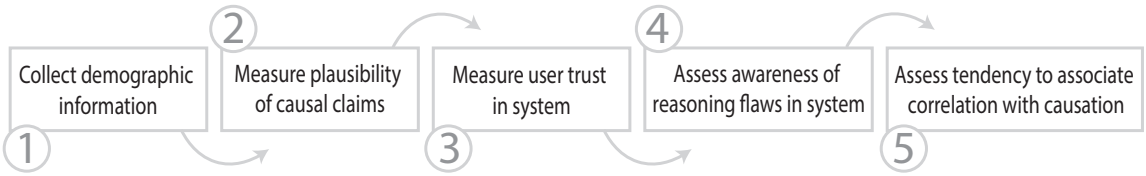


Figure 5.6: The five stages in study 1 and study 2.

In stage 1 (Figure 5.6 ①), participants filled out a demographic survey. After filling out the survey, they completed a practice task to get acquainted with the study interface.

In stage 2 (Figure 5.6 ②), participants reviewed a series of nine answers to why questions. In each task, they examined a data observation (e.g., Mississippi has the highest poverty rate among all US states) (Figure 5.5a), a why question (e.g., why is poverty rate in Mississippi so high?) (Figure 5.5a), and the system's answer to the question (Figure 5.5b-d). Depending on the condition, participants saw a different visual design for the answers. Based on the system's answer, participants rated their agreement with a causal claim (e.g., low employment rate in Mississippi is a factor that leads to high poverty rate in Mississippi) on a 7-point Likert scale.

I constructed the nine answers using the top nine causal claims obtained from the pre-study (Figure 5.4). Hence, three answers were reasonable, three were unreasonable, and three had plausibility that was difficult to judge. This intended to mirror real-world systems that tend to be unreliable in answering why questions. The order of the answers was randomized to prevent order effects.

After participants reviewed the nine answers, I measured user trust in the system in stage 3 (Figure 5.6 ③). Participants rated their trust in the system on a 7-point scale from -3 (I don't trust it at all) to +3 (I fully trust it). They further shared their reasons for trusting or not trusting the system.

Next, I assessed their awareness of the reasoning flaws in the system in stage 4 (Figure 5.6 ④). Participants reported whether they observed any reasoning flaws in the system. If the answer was "yes," I asked them to specify the reasoning flaw(s) they found.

In the final stage (Figure 5.6 ⑤), I assessed their tendency to associate correlation with causation. Participants in stage 5 saw a description of a data observation (Mississippi has the second highest child mortality among all US states) and a scatterplot showing a strong correlation between child mortality and an unknown variable X (Figure 5.7a). To assess participants' understanding of scatterplots, I first asked participants to answer two interpretation check questions (Figure 5.7b). Participants who failed any of the questions were excluded from the data analysis.

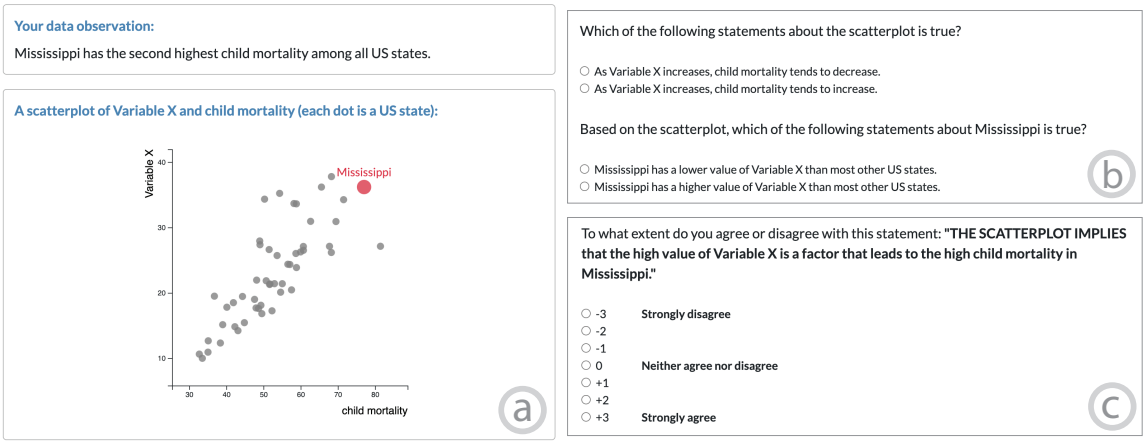


Figure 5.7: Measuring tendency to associate correlation with causation. Participants saw a data observation and a scatterplot (a), answered interpretation check questions (b), and rated their agreement on a statement suggesting that the scatterplot implied causation (c).

Whereas participants rated their agreement with a causal relationship in stage 1, participants rated their agreement with a sentence stating that a scatterplot with a high correlation implied a causal relationship in stage 5. Participants saw a statement: "*The scatterplot*

implies that the high value of Variable X is a factor that leads to the high child mortality in Mississippi" (Figure 5.7c). They rated the statement on a 7-point Likert scale and explained why they agreed or disagreed.

Quantitative Measures

I derived six measures from participants' responses.

Agreement (reasonable). For each participant, I computed the average agreement rating for the three reasonable answers.

Agreement (unreasonable). It is the average rating for the three unreasonable answers.

Agreement (hard to tell). It is the average rating for the three answers that were hard to tell if they were reasonable.

Trust. Some researchers have developed questionnaires to assess user trust in recommender systems [117] and machine learning systems [24]. However, these questionnaires may not be applicable to auto-insight systems that can answer questions about the data. I tailored a question for assessing trust in question-answering systems. In the post-study survey, I asked, "Overall, how much do you trust or not trust the question-answering system?" and participants rated on a scale from -3 to +3.

Awareness of system's flaws. I computed the number of participants who selected "yes" for the question, "Did you observe any flaw(s) in the reasoning of the question-answering system?" Unlike the other measures that are scales between -3 and +3, this measure is a count between zero and 50. Whereas trust and agreement with answers are more subjective, observations about reasoning flaws in the system are more clear-cut, making a yes/no question more suitable.

Awareness of "correlation is not causation". In the last part, participants rated a statement: "*The scatterplot implies* that the high value of Variable X is a factor that leads to the high child mortality in Mississippi." (Figure 5.7c) If participants were cautious about drawing causal conclusions from correlation, they should be inclined to disagree with the

statement.

During a pilot study, I observed that when the variable name was shown, participants tended to use their common sense to decide if they agreed with the statement. Yet, I wanted to assess tendency to confuse correlation and causation instead of ability to apply common sense. To reduce the impact of common sense in answering the question, I hid the variable name of X.

Likert-scale data are not continuous and violate the ANOVA assumptions. To study the main effect of answer design, I used a Kruskal-Wallis test, which is a non-parametric equivalence of one-way ANOVA, for the five measures using a 7-point scale (i.e., all measures except the awareness of system's flaws). When there is a significant main effect, I conducted post-hoc Wilcoxon rank sum tests with a Holm-Bonferroni correction for pairwise comparisons.

For the awareness of the system's flaws, I used a Fisher's exact test to assess if the number of participants who found reasoning flaws in the system was significantly different across conditions.

Qualitative Response

There were three open-ended questions in the survey, one for explaining trust or distrust, one for explaining reasoning flaws in the system, and one for explaining why participants agreed or disagreed that the scatterplot implied a causal conclusion.

For each question, I open-coded the responses to identify the emergent categories and develop a codebook. I observed that a response could include multiple categories. Hence, I treated each category as binary: For each response, I labelled whether each category was present or absent. A colleague and I then independently coded all responses. We then discussed inconsistencies, refined code definitions, and independently re-coded the responses based on the new definitions. We iteratively coded the responses until we reached a Cohen's κ above 0.7 for all the categories.

For each category, I conducted a Fisher's exact test to determine whether its presence is significantly different across conditions.

Hypotheses

I developed hypotheses based on research in visualization's persuasive power, trust in automated systems, and warning science.

Pandey et al. [110] found that when participants did not have a strong attitude towards a topic, visualizations had a strong power to change their attitudes. They also commented on the difficulty to change attitudes for topics of which participants already had a strong prior opinion [110]. I expected that showing a scatterplot would increase the plausibility of hard-to-tell claims because participants likely did not have a strong attitude towards them. I also expected that the scatterplot would not affect the plausibility of reasonable and unreasonable claims.

H1.1: Participants' agreement with the reasonable claims does not differ across conditions.

H1.2: Participants' agreement with the unreasonable claims does not differ across conditions.

H1.3: Participants in the three conditions that show a scatterplot in the answers (i.e., claim + vis, claim + vis + description, and claim + vis + description + warning) agree with the hard-to-tell claims more than participants in the claim-only condition.

Transparency in automated systems can inspire user trust [134]. For example, when a recommender system provides reasons behind its recommendations, users tend to trust the system more [57]. Showing the data could increase the transparency in the auto-insight system. I posited that users would trust the system more when it showed the scatterplot.

H1.4: Participants in the three conditions that show a scatterplot in the answers trust the system more than participants in the claim-only condition.

Some researchers in warning science have compared the effectiveness of passive and active warnings [41]. Whereas active warning forces users to notice it by blocking user

tasks, passive warning (e.g., a simple warning message) is less interrupting [41]. In data analysis, passive warning is more suitable because a small latency in interaction can hamper analysis quality [92]. However, Egelman et al. [41] showed that passive warnings were often ineffective because users might ignore them. The ineffectiveness might extend to auto-insight systems. Hence, I posited that the warning message would not increase participants' awareness of the system's flaws nor decrease their tendency to associate correlation with causation.

H1.5: Participants' awareness of the system's flaws does not differ across conditions.

H1.6: Participants' awareness of "correlation is not causation" does not differ across conditions.

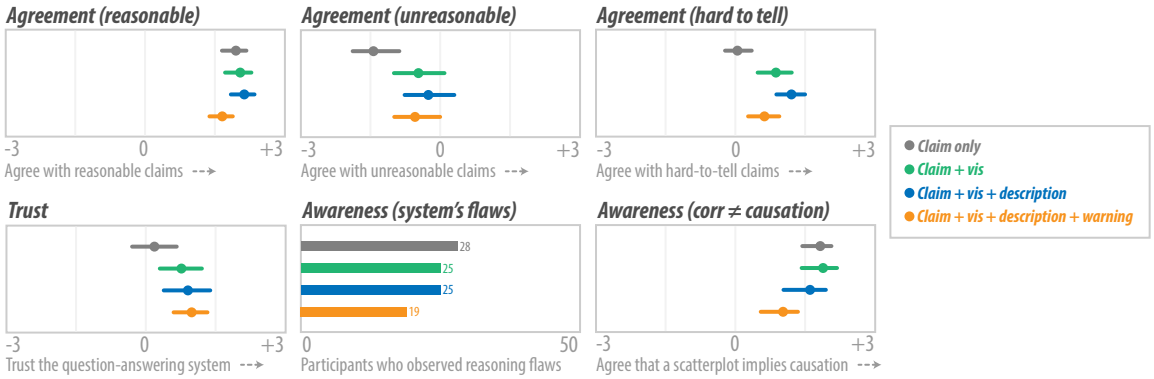


Figure 5.8: Quantitative results from study 1. All error bars show 95% bootstrapped confidence intervals.

5.4.2 Results

Figure 5.8 summarizes the results for the quantitative measures. I observed that the scatterplot increased the plausibility of unreasonable and hard-to-tell claims but not reasonable claims. Yet, textual description about correlation did not appear to affect the plausibility of causal claims. This provided evidence that by providing visual correlational evidence (e.g., a scatterplot) next to a causal claim, auto-insight systems like Explain Data [145] could leverage the persuasive power of visualizations to create an illusion of causality.

I found that the warning message appeared to decrease the plausibility of reasonable

claims but not unreasonable and hard-to-tell claims. Furthermore, the warning seemed to increase the awareness of "correlation is not causation." Despite the statistical significance, the mean rating for the statement about a scatterplot implied causation was positive for the condition with a simple warning (M=1.02). In other words, participants tended to agree that a scatterplot showing correlation in the data implied causation even when the system warned users about "correlation is not causation." This potentially indicates the ineffectiveness of a simple warning in reducing the illusion of causality.

Also, for the claim-only condition, participants tended to give a neutral rating for the hard-to-tell claims (M=0.04), a positive rating for the reasonable claims (M=1.95), and a negative rating for the unreasonable claims (M=-1.44). This confirmed the validity of the pre-study results.

In the following, I provide the detailed analysis.

Agreement (Reasonable)

On a scale from -3 (strongly disagree) to +3 (strongly agree), participants in the claim + vis + description condition rated the reasonable claims the highest (M=2.13, SD=0.92), followed by those in the claim + vis condition (M=2.04, SD=0.98), the claim-only condition (M=1.95, SD=0.96), and the claim + vis + description + warning condition (M=1.65, SD=0.89). A Kruskal-Wallis test indicated a significant main effect of answer design on the rating ($\chi^2(3)$=10.8, p=.013). I conducted six post-hoc pairwise comparisons using Wilcoxon rank sum tests with a Holm-Bonferroni correction. Results showed that only the difference between claim + vis + description + warning and claim + vis + description (p=.017) as well as that between claim + vis + description + warning and claim + vis (p=.044) were significant. The results did not support **H1.1**.

Agreement (Unreasonable)

Participants in the claim + vis + description condition rated the unreasonable claims the highest (*M*=-0.26, *SD*=1.92), followed by those in the claim + vis condition (*M*=-0.47, *SD*=1.96), the claim + vis + description + warning condition (*M*=-0.55, *SD*=1.82), and finally the claim-only condition (*M*=-1.44, *SD*=1.74). A Kruskal-Wallis test indicated a significant main effect of answer design on the rating ($\chi^2(3)$=12.2, *p*=.007), with post-hoc pairwise comparisons showing that all the three conditions with scatterplots in the answers had a significantly higher average rating than the claim-only condition. The results did not support **H1.2**.

Agreement (Hard to Tell)

Participants in the claim + vis + description condition rated the hard-to-tell claims the highest (*M*=1.2, *SD*=1.12), followed by those in the claim + vis condition (*M*=0.87, *SD*=1.34), the claim + vis + description + warning condition (*M*=0.62, *SD*=1.25), and the claim-only condition (M=0.04, SD=1.05). There is a significant main effect of answer design on the rating ($\chi^2(3)$=24.4, $p<.001$). Pairwise comparisons showed that all the three conditions with scatterplots in the answers had a significantly higher average rating than the claim-only condition. Other pairs were not significantly different. The findings supported **H1.3**.

Trust

On average, the trust ratings across conditions were positive, indicating a tendency to trust the system. Claim + vis + description + warning has the highest rating (*M*=0.98, *SD*=1.31), followed by claim + vis + description (*M*=0.9, *SD*=1.76), claim + vis (*M*=0.76, *SD*=1.67), and claim-only (*M*=0.18, *SD*=1.70). However, I did not observe a significant main effect of answer design on trust ($\chi^2(3)$=7.34, *p*=.062). The results did not support **H1.4**.

Why Trust or Not Trust?

I coded participants' reasons for trusting or not trusting the auto-insight system. Seven categories of responses emerged from the analysis. I report the core results here and provide the detailed breakdown of the categories across conditions in Appendix D.

For each response, I labelled each category as present or absent. I labelled all categories as absent for responses that were too broad or vague (e.g., *"it is nice"*).

The top three reasons for distrusting the system were some answers did not make sense (40.5% of 200), the system confused correlation and causation (9%), and it did not provide enough support for its causal claims (7.5%). A participant felt that some claims lacked support and wrote, *"Some of the answers could be factual but it was hard to determine without further data."*

The top three reasons for trusting the system were that some answers made sense (26%), the system showed the data (8.5%), and the system provided some support for its causal claims (6.5%)

I did not observe a significant difference in the presence of any of the categories across conditions using Fisher's exact tests (details in Appendix D).

Awareness of System's Flaws

Using a Fisher's exact test, I did not find a significant difference in the number of people who found reasoning flaws (these participants selected "yes" for the question asking whether they observed reasoning flaws) across conditions (p=.33). I could not reject **H1.5**.

What Are the Reasoning Flaws?

The qualitative coding resulted in four categories. Among the 97 participants who observed reasoning flaws in the system, the majority of participants stated providing nonsensical answers as a reasoning flaw (70.1% of 97). Other observed reasoning flaws were confusing correlation and causation (15.5%), not having enough support for the claims (8.25%), and

considering only one factor (3.09%). Fisher's exact tests did not indicate significant differences in the presence of any of the four categories across conditions.

Awareness of "Correlation Is Not Causation"

I asked participants to rate a sentence stating that a scatterplot implied causation. On a scale from -3 (strongly disagree) to +3 (strongly agree), claim + vis + description + warning has the lowest rating (M=1.02, SD=1.45), followed by claim + vis + description (M=1.6, SD=1.59), claim-only (M=1.82, SD=1.10), and claim + vis (M=1.88, SD=1.33). All conditions got a positive average rating, indicating a tendency to associate correlation with causation. I found a significant main effect of answer design on the rating ($\chi^2(3)$=15.2, p=.002). Post-hoc pairwise comparisons showed that claim + vis + description + warning had a significantly lower average rating than all the other three conditions, indicating that the warning appeared to reduce the tendency to associate correlation with causation. The results did not support **H1.6**.

Why Agree or Disagree With the Statement?

The qualitative coding yielded four categories. I again observed that some responses were overly broad (e.g., *"because the graph shows it"*) and coded all categories as absent for such responses.

Among the more specific responses, the majority of participants agreed that the scatterplot implied a causal relationship because the scatterplot showed a correlation (46% of 200). An example response is *"If Variable X did not rise then child mortality would not rise."*

Participants disagreed with the statement because correlation is not causation (8.5%), variable X was unknown and they could not judge (8.5%), and the scatterplot had outliers (4%). A participant who observed outliers said, *"I only slightly agree because other states show otherwise. Texas, for instance, has a much lower Child Mortality rate but Variable X*

is almost the same."

I did not find significant differences in the presence of any of the categories across conditions.

5.5 Study 2: Providing Only Reasonable Answers

Several findings from study 1 deviated from my expectations: The simple warning appeared to decrease the plausibility of reasonable claims and increase the awareness of "correlation is not causation"; I did not have enough evidence that user trust was improved by showing the data. A potential explanation lied in the unreliable performance of the system—it made the warning more noticeable and reduced the effectiveness of showing the data in improving user trust (when the system performed poorly, it is untrustworthy no matter whether it showed the data). To investigate whether the observations in study 1 held for a system that had a higher perceived performance, I conducted study 2.

5.5.1 Methods

Study 2 was the same as the study 1 except that participants reviewed nine reasonable answers to why questions (as opposed to reviewing answers with different levels of plausibility in study 1). I constructed the answers using the top nine claims in the ranked list of 30 reasonable claims obtained from the pre-study.

I similarly recruited 50 participants per condition (200 unique workers in total). Workers who participated in the pre-study and study 1 were excluded from study 2. Participants aged 18-70 (M=36.1, SD=11.0). 121 were male, 77 were female, and 2 preferred not to say. The reported educational attainments were high school (28 participants), professional school (10), college (116), graduate school (37), PhD (8), and postdoctoral (1). Concerning data analysis expertise, 44 had none, 79 were beginners, 54 were intermediate, and 23 were advanced. For experience with visualization platforms (e.g., Tableau), 84 had none, 47 were beginners, 45 were intermediate, and 24 were advanced. When asked about the

frequency of using question-answering systems, 123 reported never, 35 reported rarely, 31 reported weekly, and 11 reported daily.

As the system only presented reasonable answers, study 2 only had four measures: agreement (reasonable), trust, awareness of system's flaws, and awareness of "correlation is not causation."

In study 1, participants heeded the warning, causing them to agree less with reasonable claims and be less likely to associate correlation with causation. I expected that both effects would disappear when the system was more trustworthy. Furthermore, in study 1, showing the data using a scatterplot did not seem to improve user trust in the system. I posited that when the system provided only reasonable answers, showing the data would improve user trust in the system. I considered the same set of hypotheses as in study 1:

H2.1: Participants' agreement with the reasonable claims does not differ across conditions.

H2.2: Participants in the three conditions that show a scatterplot in the answers trust the system more than participants in the claim-only condition.

H2.3: Participants' awareness of the system's flaws does not differ across conditions.

H2.4: Participants' awareness of "correlation is not causation" does not differ across conditions.

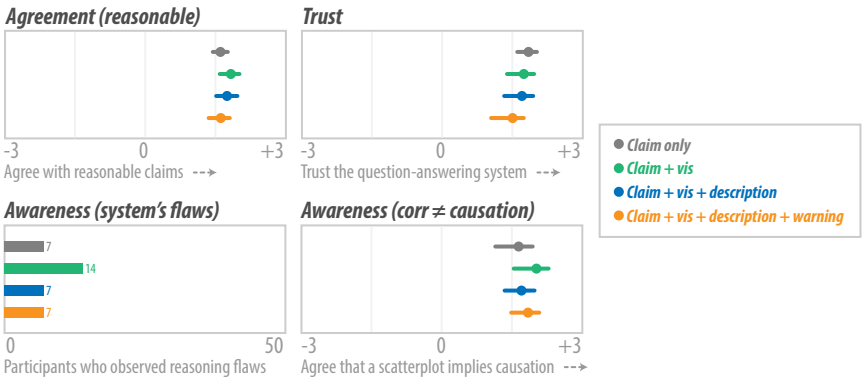


Figure 5.9: Quantitative results from study 2. All error bars show 95% bootstrapped confidence intervals.

5.5.2 Results

Figure 5.9 shows the quantitative results. Kruskal-Wallis tests for agreement (reasonable), trust, and awareness of "correlation is not causation" as well as a Fisher's exact test for awareness of system's flaws indicated no significant differences across conditions (details provided in Appendix D). Hence, the results failed to support **H2.2**. However, I could not reject **H2.1**, **H2.3**, and **H2.4**.

I also observed that participants in study 2 appeared to trust the system more than those in study 1. The mean trust rating in study 2 was 1.70 (*SD*=1.04) while that in study 1 was 0.71 (*SD*=1.64). Participants in study 2 also found fewer reasoning flaws in the system. The total number of participants who found reasoning flaws in study 2 is 35 (compared with 97 in study 1). I summarize the qualitative results as follows.

Why Trust or Not Trust?

Participants provided diverse reasons for trusting or not trusting the system. Seven categories of reasons emerged from the qualitative coding.

The top three reasons for trusting the system were the answers made sense (38.5% of 200), the system provided enough support for its causal claims (19.5%), and it showed the data (8%). The top three reasons for distrusting the system were the system did not provide enough support for its claims (8%), it considered only one factor (7%), and it confused correlation and causation (2.5%).

Fisher's exact tests indicated that the number of participants stating *the answers made sense* as a reason was significantly different across conditions (p=.014). I conducted six post-hoc pairwise comparisons using Fisher's exact tests with a Holm-Bonferroni correction. I only observed that more participants in the claim-only condition stated *the answers made sense* than in the claim + vis + description condition (p=.025). A potential explanation was that providing a claim only led participants to comment mostly on the plausibility of the claim. However, providing other information (e.g., a scatterplot) alongside a claim

enabled them to comment on other aspects and less on the plausibility.

In study 2, 56% of the responses contained reasons for trusting the system while 16% contained reasons for not trusting it. The data stood in contrast to those in study 1. In study 1, 39.5% of the responses had reasons for trust while 49.5% had reasons for distrust. This echoed the finding that participants trusted the system more in study 2.

What Are the Reasoning Flaws?

Among the 35 participants who answered "yes" for the question asking whether they observed reasoning flaws in the system, I found three categories of responses after omitting those who provided vague answers: The system considered only one factor (28.6% of 35); it confused correlation and causation (20%); it did not provide enough support for the claims (17.1%). Using Fisher's exact tests, I did not observe significant differences in the presence of the categories across conditions.

Why Agree or Disagree With the Statement?

The qualitative analysis resulted in five categories of responses. Congruent with study 1's results, most participants agreed that the scatterplot implied a causal relationship because the scatterplot showed a correlation (43.5% of 200).

Participants who disagreed with the statement commented that correlation is not causation (7.5%), the scatterplot had outliers (7%), variable X was unknown, and they could not judge (5.5%), and the dots in the scatterplot looked disperse (1.5%).

Fisher's exact tests did not show a significant difference across conditions for any of the categories.

5.6 Discussion

Before discussing the implications of the findings, I summarize the results from the two experiments and offer potential explanations for the less intuitive observations.

In study 1, participants reviewed answers of different plausibility. I did not observe an effect of the textual description about correlation on the plausibility of causal claims, user trust, awareness of the system's flaws, and awareness of "correlation is not causation." However, showing a scatterplot caused participants to disagree less with unreasonable claims and agree more with hard-to-tell claims. This implies that auto-insight systems could mislead users by visualizing correlation in the data to induce an illusion of causality.

Furthermore, in study 1, the warning message seemed to cause participants to agree less with reasonable claims. The warning also reduced participants' tendency to associate correlation with causation. However, the effects appeared to be unreliable. When participants examined only reasonable answers in study 2, I could not observe the effects of the warning message on reducing the plausibility of reasonable claims and on raising the awareness of "correlation is not causation." Moreover, in both studies, I found that the ratings for the awareness of "correlation is not causation" were positive (i.e., agreeing that the scatterplot implied causation) even when participants were warned that correlation is not causation. This implies that although a simple warning message could be useful in promoting a correct interpretation of auto-insights about causation, it may not be enough, and researchers could explore alternative design ideas.

Why did the effects of the warning disappear in study 2? Research in warning science found that arousal strength (i.e., the perceived importance or relevance of a warning) affects the effectiveness of a warning message in motivating safety-related behaviors [56]. Participants in study 2 tended to trust the system more than those in study 1. This likely led participants in study 2 to perceive the warning about the system's reasoning flaws to be less relevant. The warning in study 2 became less effective possibly because participants tended to ignore the warning.

In both studies, I did not observe significant differences in user trust and the awareness of the system's flaws across conditions. The qualitative results provided an explanation. In study 1, when asked about why they did not trust the system or what were the reasoning

flaws in the system, most participants simply stated that the answers did not make sense. In study 2, when asked about why they trusted the system, the majority commented that the answers made sense to them. The results appeared to indicate that system performance in answering why questions had a dominating effect on user trust and the awareness of reasoning flaws in the system. In other words, when users can assess system performance, showing other information (e.g., a scatterplot or a warning) may play a small role in shaping user trust and the awareness of flaws.

I further observed a general tendency for participants to conclude causation from correlation. Overall, participants tended to agree that a scatterplot that visualized correlation between variable implied a causal relationship. How do we reduce the illusion of causality when using auto-insight systems like Explain Data [145]? Here, I devise design considerations based on the study results.

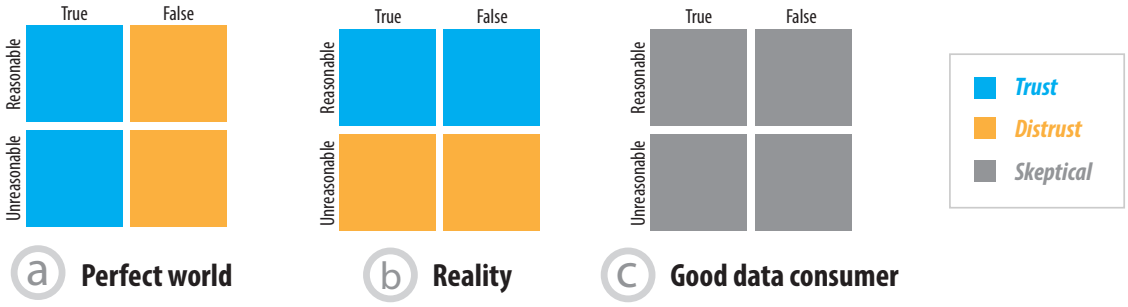


Figure 5.10: The relationship among trust, a claim's plausibility and the ground truth in different scenarios. In each bigger square, the y-axis is a claim's plausibility and the x-axis is the ground truth. In a perfect world (a), users should trust a claim only if it is true. In reality (b), users tend to trust a reasonable claim and distrust an unreasonable claim. A good data consumer (c) should be skeptical despite a claims' plausibility when the truth is unknown.

5.6.1 Inspiring Skepticism in Auto-Insights About Causation

A core implication of the results is that auto-insights systems could utilize visual correlational evidence to create an illusion of causality: By showing a scatterplot, these systems could increase the likelihood for users to accept causal claims that might be unfounded.

Mitigating the misleading power of auto-insights about causation entails a deliberate design effort. In this section, I advocate that system designers could encourage users to be skeptical about auto-insights that describe causation and propose design ideas to inspire skepticism.

Why should users be skeptical when considering causal claims that are automatically generated? In a perfect world, user trust in a causal claim should match the ground truth—users should trust causal claims only when they are true and distrust false claims (Figure 5.10a). In reality, however, belief in causal claims depends on their perceived plausibility despite the ground truth (Figure 5.10b). Very often, users can only determine the plausibility of a causal claim but not whether it is true. Hence, I argue that users should be skeptical whenever they cannot assess the veracity of a causal claim (Figure 5.10c): A good data consumer should question the validity of a reasonable claim because the causal relationship could be fake; she should not refute the possibility of an unreasonable claim since the claim could hold true.

Encouraging Skepticism for Reasonable Claims

How do we encourage users to be skeptical about reasonable causal claims through interface design? Warning could be a potential solution. In study 1, I observed that participants tended to be more cautious in agreeing with a reasonable claim given a warning message. Yet, the effect of a simple warning message appeared to be unreliable: It disappeared in study 2 in which the system only provided reasonable causal claims. To improve the effectiveness of warning in inspiring skepticism, it could be designed based on research in warning science. For example, Wogalter [162] proposed the Communication-Human Information Processing (C-HIP) Model to describe the perceptual and cognitive processes after people see a warning. The model suggests asking a series of questions to assess the effectiveness of a warning. For instance, do people notice the warning? Is the message in the warning being effectively communicated?

What are alternative interface design solutions to help encourage skepticism? In study 1, I found that participants tended to disagree less with unreasonable claims and agree more with hard-to-tell claims given a scatterplot. The finding reveals that users might believe in a false causal relationship based only on evidence about correlation in a scatterplot. To mitigate the misleading power, it seems plausible to hide scatterplots from causal claims. Yet, some participants felt that they trusted the system more because the scatterplots enabled them to see the data. An ideal solution is to keep the benefits of scatterplots while mitigating their potential power to mislead users. Ritchie et al. [122] found that transitioning from a non-deceptive view to a deceptive one could reduce the deception caused by the second view while enabling users to access the benefits of this view. Similarly, an auto-insight system could hide scatterplots by default while providing an option for users to view them. It would be interesting to investigate whether this design can reduce users' tendency to confuse correlation and causation.

How do we improve general awareness of "correlation is not causation"? Besides inspiring skepticism about a reasonable claim, warning also appeared to raise awareness of "correlation is not causation" in study 1. Again, the effect of a simple warning message could be weak. Indeed, participants tended to agree that the scatterplot implied causation even in the condition with the warning message. This suggests that system developers might need to look beyond interface design to help users acquire correct statistical knowledge. Alternatives include pedagogical approaches such as tutorials. For example, when Tableau introduced Explain Data, they emphasized that users were the data experts, and they should judge the validity of the causal claims based on their knowledge [143, 144]. Future work will study the effectiveness of such tutorials on reducing users' tendency to associate correlation with causation.

Encouraging Open-Mindedness for Unreasonable Claims

While this work sheds light on ways to inspire skepticism for reasonable claims, designs to keep users open-minded when they see unreasonable claims are yet to be explored. Open-mindedness is a different form of skepticism: Instead of being skeptical about the automatically generated causal claims, users are skeptical about their beliefs, hypotheses, and expectations about the data. Different models (e.g., Bayesian statistics [10] and the data-frame model [72, 73]) have been developed to explain the process through which people update their beliefs. Prior research in misinformation showed that existing beliefs are rigid and people are inclined to resist changes to their beliefs [86].

While encouraging open-mindedness could be challenging, what are some potential ideas to keep people open-minded when using auto-insight systems like Explain Data [145]? An idea is to enable users to tell the system if an answer makes sense. If users consider an answer nonsensical, the system could explain why a causal relationship might exist to prevent users from prematurely rejecting a causal claim that seems unreasonable. However, further evidence would be required to demonstrate the effectiveness of this approach.

5.6.2 Study Limitations and Future Work

Results from the studies hint at the potential for scatterplots to create an illusion of causality and the potential for a simple warning to reduce this illusion. I note that these are observations under controlled experiments, and I am prudent in drawing conclusions about the practical significance of the findings. First, to collect data from hundreds of people on MTurk, I needed to sacrifice realism to adapt the study for an online setting. For example, participants examined a series of answers provided by the system rather than really interact with a working system. Second, collaboration could protect users from being misled in practice: While an analyst might draw causal conclusions from correlational evidence, colleagues might remind the analyst of the flaw. Learning about the practical implications of my findings will require observing how people employ systems such as Explain Data [145]

in their workflow and studying how people collaborate during data analysis.

I also note that the effectiveness of a simple warning in reducing causal illusion warrants further studies. My studies only compared four experimental conditions (Figure 5.5). However, evidence from further comparisons (e.g., a comparison between an additional claim + vis + warning condition and the original claim + vis condition) could support the effectiveness (or ineffectiveness) of the warning. An ideal experiment is to consider each information type (claim, vis, textual description, and warning) an independent variable with two levels (with and without). This experiment will enable comparisons among all possible experimental conditions. Nevertheless, adding more conditions greatly reduces power given Bonferroni correction, and interesting findings would be missed. In future studies, experimenters would likely want to preserve power by honing in on a smaller set of comparisons. My findings could provide guidance on what focused comparisons to make.

My target population was potential end users of auto-insight systems. These users include both people who are less proficient in data analysis and those who are more proficient. My participants ranged from beginner users to more advanced analysts and appeared to be a reasonable proxy for my target. Yet, the focus on these users also implies that some findings (e.g., the tendency among participants to confuse correlation and causation) may not generalize if I conduct the studies with professional analysts only. Future work will replicate my study with these experts.

In both studies, I used a single question to measure trust in the auto-insight system. In future, a questionnaire with multiple questions could be developed for assessing user trust in these systems. Such a questionnaire will measure sub-dimensions of trust (e.g., understanding) and enable researchers to learn about more fine-grained reasons for trusting a system (e.g., the system is trustworthy because users can easily understand the answers).

I have investigated whether showing correlational evidence could induce causal illusion. One form of correlational evidence I studied was textual description about correlation (e.g., as employment rate decreases, poverty rate tends to increase). Designing a description that

completely eliminates casual perception could be challenging because people might easily mistake correlation for causation. Future research will investigate how different phrasing of correlation descriptions will affect casual perception.

My study focused on numerical variables. It is common to visualize correlation between numerical variables using scatterplots. For other data types (e.g., categorical variables), alternative visualizations are possible. A natural extension to this work is to investigate the generalizability of my findings to other data types and visualizations.

Finally, although I am an advocate for encouraging skepticism when using auto-insight systems, I note that inspiring the right level of skepticism could be challenging. Ideally, users' skepticism about a causal claim should match the evidence they have about the claim: For causal claims with good support (e.g., carbon dioxide emission leads to global warming), users could be less skeptical; for claims that lack supportive evidence, users could evaluate them more critically. However, it is difficult for a system to infer the amount of evidence users have about a claim and encourage skepticism accordingly. Moving forward, researchers could investigate whether telling users to be skeptical (e.g., through warnings) promotes an appropriate level of skepticism or engenders excessive and unhealthy skepticism.

CHAPTER 6
DISCUSSION

This chapter discusses a vision of future auto-insight systems, the remaining challenges, and the directions for future research. At the end, I conclude with the broader implications of my dissertation.

6.1 The Future of Automated Insight

Not only could future auto-insight systems automatically reveal interesting data relationships, they could also understand the context of data analysis, intelligently enable scrutiny, and proactively protect users from being misled. Let us consider what such a future might look like.

Suppose that Ella is a business analyst in a tech company. Day in and day out, she analyzes data about her organization's sales and operations to glean actionable insights for the C-suite. One day, she is preparing a dashboard about the trends in sales for a presentation to her manager.

Having loaded the recent sales data into an intelligent visualization platform, Ella sees recommendations about interesting patterns in sales. The system is smart. Knowing that the organization has released a new version of its major product, it prioritizes information about this product. Nonintrusive recommendation ensures that Ella's flow of analysis is not interrupted. Seeing the recommendations, Ella is a bit worried—the sales of their major product have been decreasing since the last release.

Ella feels that the decreasing trend is something worth diving into. Sure enough, she does not trust the recommendation one hundred percent—she will be presenting the dashboard to her manager, and errors mean embarrassment. She asks the systems some clarifying questions, *"Is it just US data? Have you included the data for regions outside the US?*

Show me the trends for California, Georgia, and New York." With no delay, the system answers Ella's questions with additional charts. Looking at the answers, Ella learns that the decreasing trend appears across the board.

"Maybe our clients don't like the latest update?" Ella is thinking. She asks, *"Can you tell me why the sales have been decreasing?"* Based on a database of domain knowledge, the system provides explanations for the decreasing trend: *"Here are some possible explanations: We spent fewer dollars on marketing this product; the sales of a complementary product have also been dropping during the same period; and..."* Ella has heard from the C-suite that the company had shifted attention from marketing existing products to innovating new ones.

Ella wants to highlight potential actions to take in the next meeting with her manager. *"What can we do?"* Ella asks the system. This is not an easy question to answer. However, the system has data about what competitors did in a similar scenario and is ready to prescribe some actions: *"In 2030, Apple tried to..., and the sales had gone up by 5% the following year."*

Our system knows the users—they could rely on the recommended data observations, explanations, and actions. When providing these recommendations, the system reveals the uncertainty in the recommendations and warns Ella about the potential inaccuracy. Ella finds the system trustworthy: *"You are smart, dude!"* She is happy about what she learnt from the system and feels good about her presentation.

Coming back from this imaginary scenario, how do we design an auto-insight system that is capable of recommending data observations, explanations, and actions, that offers transparency in the recommendations, that respects user agency, and that proactively protects users from misinterpretation? Here, I identify five areas of challenges and future research directions.

6.1.1 Providing Transparency in Recommendations

Throughout this dissertation, I have highlighted that solely providing recommendations of statistical information is not enough for the adoption of auto-insight systems in practice. From my interactions with visualization practitioners, I have learnt that it is imperative to enable users to validate the recommended information and to ask follow-up questions about it. Such transparency in auto-insight systems could help users develop trust in the recommendations, which is an important consideration in making the recommendations useful and insightful.

In the above scenario, Ella asks clarifying questions (e.g., *"Have you included the data for regions outside the US?"*) and follow-up questions (e.g., *"Show me the trends for California, Georgia, and New York"*) to gather a multi-faceted view of a recommended data observation. The ultimate vision is an auto-insight system that can explain itself. Other fields of artificial intelligence illuminate such a possibility: Autonomous game agents can rationalize their moves in games in a way that non-experts can understand [42, 43]; when answering questions about charts, systems can generate natural language explanations that reference chart features [69]; conversational agents can explain its recommendations about movies just as humans would [113]. The possibility of explainability in auto-insight systems raises questions for future research: What are the types of explanations users want? How do these explanations shape user perceptions of the system?

The workflow of data analysis is fundamentally situated [141]—it is shaped by the confluence of organization culture, work context, and socialization. The types of explanations and the perceptions of users therefore depend on the context where the data analysis resides. Hence, answering the research questions I just described entails a better understanding of users. For example, why is transparency in auto-insight systems desirable to practitioners? My interviews point to a diverse set of reasons: Effectiveness in communicating data insights requires the ability to answer questions from the audience; soundness in decision making warrants accuracy in data insights; the truthfulness of scientific knowledge depends

on meticulous validation. This set of reasons is only a starting point for understanding user needs and concerns. Studying transparency in auto-insight systems will benefit from future interviews and close collaboration with users.

6.1.2 Having Knowledge about the Context

In this dissertation, auto-insights mostly concerns the recommendation of statistical information through visualizations and data facts [136, 156]. These recommendations include automatically identified data observations (e.g., statistical information about correlation) and explanations for the observations (e.g., statistical information about causation). Other than observations and explanations, the scenario about Ella also highlights the recommendation of possible actions (e.g., what to do given the sales are down?). The idea to recommend data observations, explanations, and actions is indeed grounded in my interviews. One of my interviewees was a president in a start-up consulting company, and here was what he said:

> *Depending on how mature an organization is analytically on data, that's going to determine a lot of the questions that they ask [...] The first question they're always going to ask is what's happening, and that's kind of descriptive [...] It's kind of like read me the news. And as soon as you understand what's happening, then it's like what's wrong, what do I need to go take action on [...] kind of a why is that happening [...] And then you start to get into what's going to happen next and what should I do about it. Kind of a postscriptive analysis rather than descriptive. When you start to be able to prescribe to the end user and guide the end user towards what they should be doing that's kind of the best outcome at the end of the day.*

How do we recommend data observations, explanations, and actions in a form users would consider relevant and insightful? I believe that domain knowledge and knowledge about the context of the data analysis is crucial.

Concerning *data observations*, my interviewees often reported that insight occurred when different information came together. I tend to agree with them. Consider this observation about rainfall: In 2020, Atlanta had 51.5 inches of rain while Seattle had 38 inches of rain [135]. You may not find this information insightful, but this is surprising to me because I have been hearing people saying that Seattle rains all the time. (Yet, the numbers show that Atlanta rained more.) This leads me to think that Seattle had a lower rainfall probably because Seattle drizzles frequently while Atlanta has occasional heavy rains. Putting different information together like the human brain does is non-trivial. Knowledge about the domain and context is often not stored in the data. Even if we have ways to represent this knowledge, automatically drawing connections among information like human intuition does could be challenging.

The recommendation of *explanations* will benefit from a knowledge base of common sense. In the previous chapters, we have seen that auto-insight systems like Explain Data [145] often provide unreasonable causal claims. If these systems have common sense, many false claims could be avoided. For instance, a system might discover a correlation between the number of smokers and the number of people with lung cancer. It would not say that a high percentage of smokers is caused by a high number of people with lung cancer if it knows the proper directionality. A challenge though is the creation of a knowledge base in a scalable manner. Conventional artificial intelligence often requires subject matter experts to manually encode their knowledge in a knowledge base [140]. Since it is impossible for a knowledge engineer to identify and encode every signal causal relationship, researchers could explore a scalable approach to mining a causal knowledge base. Fast et al. [45] demonstrated how a vector space model of human activities could be automatically generated from 1.8 billion words of modern fiction. This indicates that messy data like fiction could also be a reliable source of commonsense knowledge.

The recommendation of *actions* could be scaffolded by a knowledge base about the actions people typically take in particular scenarios. In the operations of an organization,

some actions are standard. For example, one of my interviewees worked in a large multinational beverage company, and he discussed what actions could be taken when the inventory was about to obsolete:

> *For example, my report could say in these five locations, you have a million cases that's about to expire. What are you going to do with them? Are you going to ship them to other locations that need them and tell them not to produce anymore in that location? [...] Or are you going to have a 99% sale to try to move them? Or are you going to give them away to charity?*

Some visualization researchers have also developed interfaces for conducting prescriptive analysis using event sequence data [40]. These systems could help users decide the appropriate actions by learning from history. For example, a patient suffering from a knee injury may want to choose between a surgery or a more conservative treatment [40]. They could learn about the typical treatment and recovery time from the electronic health records of patients who similarly suffered from knee injuries [40]. Future research will explore the automation of prescriptive analysis using event sequence data and investigate how users perceive the recommendation of actions.

6.1.3 Preventing Users from Being Misled

In the story about Ella, the system has a model of an imperfect user. The system knows that users could easily be misled and appropriately warns users about the potential unreliability in the recommendations. Realizing this vision requires a substantive understanding of specific types of misinterpretation. The frameworks offered in Chapter 3 could help set research priority. I have proposed 14 types of statistical information that could be automatically generated and 5 sources of unreliability in this information. One such unreliability is sampling variability, which implies that observations from a sample of data cannot be generalized to the population. Without information about the uncertainty in a data sample, users could mistake patterns observed from a sample as a phenomenon that exists in

the population. Future research in the misinterpretation of auto-insights will still surround the two questions I posed at the very beginning of this dissertation: Could auto-insights mislead users in a particular way? If they could, how to protect users from being misled?

Misinterpretation of auto-insights occurs when systems provide false claims about the data, but users believe these claims. The definition of misinformation ("false information that is disseminated, regardless of intent to mislead" [85]) aptly describes the provision of false claims by auto-insight systems. Different from conventional misinformation that is spread through news and social media, the misinformation I have studied is generated by automated systems. Nevertheless, lessons concerning ways to protect users from misinterpreting auto-insights could potentially be learnt from research in misinformation.

For example, simply warning upfront that the information people are about to see is misleading could reduce reliance on misinformation [87]. In Chapter 5, my crowdsourced studies showed that presenting a warning message next to a causal claim might fail to create an awareness of "correlation is not causation." To enhance its effectiveness, warning about the unreliability in the recommendations could be presented *before* people use an auto-insight system.

Furthermore, encouraging people to evaluate information more critically could protect them from misinformation [120]. This strategy is particularly effective when people are educated about how to be critical (e.g., slowing down when thinking about the information and considering the credibility of the information source) [85]. Hence, aside from warning, another approach to avoiding the misinterpretation of auto-insights is to encourage criticism. How do we design an interface to support critical evaluation of information? Answering this question warrants further design research and user studies.

6.1.4 Respecting User Agency

The system in our imaginary scenario respects user agency by providing recommendations in an unobtrusive manner. However, an appropriate balance between user agency and

automation is difficult to design for. For example, while being unobtrusive, the recommendations can still indulge reliance, reducing the freedom for users to take initiative in data exploration. This was indeed a concern of some of my interviewees.

Researchers have developed different frameworks for considering an appropriate balance between user agency and automation. Horvitz [59] proposed the principles of mixed-initiative user interfaces in which the automation guesses the goals and needs of users, and users guide the provision and invocation of automated services. Due to the uncertainties in user's goals, he suggested that systems could engage users in resolving the uncertainties, minimize the cost of poor guesses, and enable users to easily dismiss the automated services [59]. While these principles provide useful guidance to system developers, effort to design appropriate automation often flounders. For example, more than a decade of research has been devoted to developing automated services that interrupt users at the right time. Due to the ambiguity in user intent (e.g., whether users are available for interruption), techniques that automatically identify interruptible moments (e.g., [2, 63]) are far from perfect, and users often express frustration due to interruptions from digital devices [49].

Another framework for considering the balance between user agency and automation concerns appropriate trust. Lee and See [83] suggested that user trust should match the capability of automation: Overtrust occurs when users trust an automated system that does not perform well; distrust occurs when users fail to trust a competent automated system; trust is appropriate when users trust the automation as much as its performance. The misinterpretation of auto-insights is indeed a problem of overtrust—users put too much trust in auto-insight systems that provide unreliable claims about the data, leading to the misinterpretation. In Chapter 5, I proposed to encourage skepticism as a solution to ensuring a correct interpretation. Skepticism essentially calibrates user trust to match the unreliability of the statistical information recommended by auto-insight systems. Applying this framework, however, could be challenging in reality because it is often difficult to determine whether user trust is appropriate. What does it mean to ensure a balance between user

agency and automation? Answers to this question will still be subject to ongoing debates.

6.1.5 Facilitating Collaborative Analysis

This dissertation has only addressed the use of auto-insight systems by a single user. In practice, data analysis is a social process. While working towards a collective goal, people in a team or an organization contribute different expertise and knowledge. Socialization could present additional complexities to the study of auto-insight systems.

What are the implications of collaborative analysis for the misleading power of auto-insight systems? A lone user might be vulnerable to the misinterpretation of auto-insights. When people work together, however, they offer different perspectives to consider the data, potentially inoculating data misinterpretation as a group. For example, in the scenario about Ella, the decreasing trend in sales may be a false discovery. While Ella might have faith in the data observation, her manager might be more skeptical and recognize that the trend is not real. What are the factors that protect people from being misled in a collaborative environment? When a group of people is misled by an auto-insight system, how does the misinformation disseminate and distribute within the group? The rich literature in Computer Supported Cooperative Work [50] and distributed cognition [93] could shed light on answers to these questions.

Besides misinterpretation, a collaborative setting changes the requirements for designing systems that automate data insights. Many of my interviewees conducted data analysis and created dashboards for some subject matter experts. They often encountered a separation of domain knowledge and technical expertise in the collaboration: While the subject matter experts had domain knowledge, they were not ones who conducted the analysis; while my interviewees had significant expertise in data analysis, they might lack the domain knowledge required for the analysis. Such a separation hints at new challenges in the recommendation of data insights. For example, users of an auto-insight system may lack domain knowledge to assess whether a recommended data observation is relevant.

The subject matter experts, who may lack expertise in using the system, could better assess the relevance. To facilitate collaborative insight generation, a design idea is to enable users to bookmark the potentially useful auto-insights for review by subject matter experts (e.g., [55, 154]). Other ideas inspired by existing collaborative systems include adding tags and links to the bookmarks for a better organization [160] and embedding usage pattern information in interface widgets to guide the exploration of auto-insights [159]. At the end of the day, designing collaborative auto-insight systems requires studying the nature of collaboration and the expertise of different stakeholders.

6.2 Broader Implications

At the core, this dissertation concerns the conceptualization of human-centered artificial intelligence in data analysis systems. The heart of human-centered artificial intelligence is to develop artificial intelligence that can understand users and that helps users understand it [121]. My research in the misinterpretation of auto-insights centers on the latter: helping users develop an appropriate understanding of artificial intelligence. In this dissertation, I provided evidence from interviews and controlled experiments to illustrate that users could misinterpret the recommendations from auto-insight systems. Based on these findings, I offered considerations for designing safer artificial intelligence in data analysis systems (e.g., designing more effective warnings to avoid misinterpretation and training users to help them acquire statistical knowledge). My dissertation further pointed to promising research directions towards more intelligible and usable artificial intelligence (e.g., developing a system that can explain itself to enhance transparency and mining a knowledge base of common sense in a scalable manner). I believe there is a bright future ahead of us—one in which humans and artificial intelligence collaborate harmoniously to solve important problems. This dissertation takes a small but critical step towards this vision.

CHAPTER 7
CONCLUSION

While visualization researchers have proclaimed insight as a core purpose of data visualization [18], auto-insight systems offer a prospect to automate the production of data insights. In face of this exciting prospect, we can easily omit the potential side effects such technology might bring. The thesis of my dissertation highlights one side effect when employing auto-insight systems for data analysis: ***User misinterpretation constitutes a concern in the use of auto-insight systems, and in particular, auto-insight systems that provide causal explanations can mislead users to interpret correlation as causation.***

To provide evidence to support my thesis, I conducted an interview study and a series of crowd-sourced studies to address three research questions:

- **RQ1**: What are visualization practitioners' potential concerns about the use of auto-insight systems?
- **RQ2**: Given that some practitioners are concerned about misinterpreting auto-insights, do people misinterpret auto-insights that describe causal claims?
- **RQ3**: If people do misinterpret auto-insights that describe causal claims, does providing a simple warning message help ensure a correct interpretation?

Visualization researchers have started a conversation about the ethical implications of data visualization [29, 98]. By addressing the above research questions, my research adds to the discussions through two contributions:

- **Articulating the side effects of automating data analysis**: Findings from an interview study with 23 visualization practitioners revealed five concerns in the use of

auto-insight systems: misinterpretation, non-transparency, information overloading, loss of agency, and misguided data exploration (**RQ1**).

- **Offering evidence that auto-insights could mislead users**: Results from experiments with more than 400 crowd workers showed that auto-insight systems could mislead users to draw conclusions about causation by presenting information about correlation (**RQ2**) and that simple warning messages could be insufficient in protecting users from being misled (**RQ3**).

These findings could offer guidance to researchers and system developers who aim to design auto-insight systems that are powerful, intelligent, and safe-to-use.

Appendices

APPENDIX A
INTERVIEW SCRIPT

I am Terrance Law and I am a PhD student at the Visualization Lab at Georgia Tech.

We are conducting a study about the practice of visualization end users and their views on automated insights in visualization platforms such as Tableau and Power BI.

This interview will take less than an hour. Can I record audio?

Part 1: Background

1. Can you tell me a bit about your year of experience with Tableau and other visualization tools? What is your role in your team? How does your job relate to understanding data?

Great! Thank you!

Part 2: Current Practice

1. Can you walk me through a typical or recent data analysis scenario in which you use visualization tools to analyse or understand data?
    - Where do the data come from?
    - What is the purpose of the analysis?
2. How much visualizations are involved in the analysis? Do you do use non-visualization tools for doing the analysis?
3. Can you tell me a scenario in which you use visualization tools to report your findings?

- Who is the audience?

4. Can you tell me about the challenges you face in your analysis workflow?

    - Ask about other challenges if talk about data preparation.

Part 3: Perceptions of Auto-Insight Systems

[Explain what auto-insight systems are at a high level: E.g., recommendations of visualizations, textual descriptions of statistical information...]

Let me show you demos of three auto-insight systems. These systems are inspired by existing products and I have some questions about their utility to your workflow.

I am going to use a US college data set for all the demos. In the data set, each row is the record of a college in a year. There are attributes such as Admission Rate, SAT Average, and Average Cost of Attendance in a year.

[Order of presentation counterbalanced]

[System A: `https://terrancelaw.github.io/iQuery/`] This system is modelled on Tableau. You can drag and drop attributes to the shelves to create visualizations. While you are creating visualizations, this pane here shows you some other interesting visualizations based on the selected attributes. For example, if I select time, it shows you some trends. If I drag X and Y onto the shelves, it will show me some statistical information relating to these attributes. When you click on a recommendation, the recommendation will be shown in the main view.

[System B: `https://terrancelaw.github.io/iPage/`] This system is modelled on Microsoft Power BI. As you load your data into the system, it shows you a page of findings before you even start your data analysis. For example, is tells you things like [...] As you review this page of recommendations, you may find something interesting, so you can bookmark the recommendations. The bookmarked recommendations will appear in a dashboard. You can edit the dashboard by moving the charts around and adding some

annotations.

[System C: `https://terrancelaw.github.io/iDashboard/`] This system is inspired by a product of a company called Narrative Science. Here is a dashboard on the right. On the left are some computationally generated findings from the dashboard. For example, from this view, the system finds that [...] From this view, the system finds that [...] As you change the attribute in a view, the findings are updated. You can also click on a state to update the text. For example, as you click on California, the text will be updated to tell me things relating to California.

1. Can you see yourself using this system to help with your workflow? If no, what do you think might be the limitations of the system? What are your concerns about using it? If yes, how do you think the system will help?

2. Ask about limitations and concerns if say something positive for the first one.

3. What are the kinds of auto-insight systems you want to have for your workflow?

Part 4: Experiences with Data Insights

1. We want to get a sense of what are the findings that people think are insightful. I wonder if you could tell me a finding, maybe from your data, that you think is really insightful.

    - Why is that an insight to you?
    - What are the characteristics of the finding that makes it insightful?

2. I wonder if you had any eureka moments in your data analysis. If yes, can you share a bit with me?

Thank you very much for you time!

APPENDIX B

RANKED LISTS OF 90 CAUSAL CLAIMS

R: Number of "Reasonable" votes

U: Number of "Unreasonable" votes

N: Number of "Not Sure" votes

B.1 Ranked List of 30 Reasonable Claims (In Ascending Order of Entropy)

	R	U	N
1. Low employment rate may be a factor that leads to high poverty rate.	20	0	0
2. High access to exercise opportunities may be a factor that leads to high percentage of adults with excellent health.	20	0	0
3. High percentage of teens who abuse alcohol may be a factor that leads to high percentage of driving deaths with alcohol involvement.	19	1	0
4. High percentage of adults who have depression may be a factor that leads to high percentage of adults with serious thoughts of suicide.	18	2	0
5. High percentage of adults who do not do exercise may be a factor that leads to high percentage of people who have poor health.	18	0	2
6. High percentage of adults with hypertension may be a factor that leads to high percentage of adults with cardiovascular disease.	18	0	2
7. High percentage of adults with diabetes may be a factor that leads to high percentage of people who have poor health.	18	1	1
8. High percentage of adults who smoke may be a factor that leads to high death rate per 100,000 population.	17	3	0
9. High percentage of preterm births may be a factor that leads to high infant death rate.	17	3	0
10. High percentage of adults with hypertension may be a factor that leads to high percentage of people who have poor health.	16	4	0

11. Low employment rate may be a factor that leads to high percentage of people who are extremely poor.	17	2	1
12. High percentage of adults who are obese may be a factor that leads to low percentage of adults with excellent health.	15	5	0
13. Low percentage of adults who exercise regularly may be a factor that leads to low average life expectancy.	14	6	0
14. High percentage of adults with at least college education may be a factor that leads to high percentage of adults who are professionals or managers.	16	3	1
15. High percentage of people who are extremely poor may be a factor that leads to high percentage of people who lack adequate access to food.	16	3	1
16. High percentage of adults who exercise regularly may be a factor that leads to low percentage of adults with cardiovascular disease.	16	1	3
17. High poverty rate may be a factor that leads to low percentage of households with internet access.	13	7	0
18. High percentage of people who are Hispanic or Latino may be a factor that leads to high percentage of people who speak Spanish at home.	15	1	4
19. High percentage of people who have poor health may be a factor that leads to low average life expectancy.	15	3	2
20. High percentage of children in single-parent households may be a factor that leads to high percentage of children in poverty.	15	3	2
21. High percentage of men who visited a dentist regularly may be a factor that leads to low percentage of adults age 65+ with all teeth extracted.	15	3	2
22. High percentage of people who are extremely poor may be a factor that leads to low percent of households with computer and smartphone.	14	5	1
23. Low percentage of adults who exercise regularly may be a factor that leads to high percentage of adults with high cholesterol.	14	4	2

24. High percentage of people born in a foreign country may be a factor that leads to high percentage of people who don't speak English at home.	12	7	1
25. High percentage of people with vision difficulty may be a factor that leads to high percentage of people with independent-living difficulty.	9	10	1
26. Low poverty rate may be a factor that leads to low percentage of children eligible for reduced price lunch.	13	4	3
27. High median household income may be a factor that leads to high percentage of adults with excellent health.	12	5	3
28. Low percentage of adults with excellent health may be a factor that leads to high percentage of adults with cardiovascular disease.	6	11	3
29. High percentage of adults who smoke may be a factor that leads to high percentage of adults age 65+ with all teeth extracted.	9	6	5
30. Low percentage of adults with at least college education may be a factor that leads to high number of people who are in jail (per 100,000 residents).	5	9	6

B.2 Ranked List of 30 Unreasonable Claims (In Ascending Order of Entropy)

	R	U	N
1. Low percentage of adults who are obese may be a factor that leads to high median housing price.	1	19	0
2. High percentage of adults age 65+ with all teeth extracted may be a factor that leads to low percentage of households with internet access.	2	18	0
3. Low motor vehicle crash deaths may be a factor that leads to high average ACT Math score.	2	18	0
4. High percentage of men who visited a dentist regularly may be a factor that leads to low rate of teenage pregnancy.	2	18	0
5. High percentage of adults age 65+ with all teeth extracted may be a factor that leads to low access to exercise opportunities.	2	18	0
6. High percentage of adults age 65+ with all teeth extracted may be a factor that leads to low median household income.	3	17	0

7. High percentage of men who visited a dentist regularly may be a factor that leads to low percentage of men who cannot afford to see a doctor.	2	17	1
8. High percentage of adults who smoke may be a factor that leads to high percentage of people with high-school education only.	5	15	0
9. Low percentage of women who smoke during pregnancy may be a factor that leads to low percentage of people who speak English only at home.	1	16	3
10. High percentage of men who visited a dentist regularly may be a factor that leads to low motor vehicle crash deaths.	1	16	3
11. High percentage of adults with hypertension may be a factor that leads to low percentage of adults with at least college education.	3	16	1
12. High access to dental care may be a factor that leads to low motor vehicle crash deaths.	3	16	1
13. Low percentage of households with internet access may be a factor that leads to high percentage of people with vision difficulty.	3	16	1
14. Low motor vehicle crash deaths may be a factor that leads to high median household income.	2	16	2
15. Low percentage of women who smoke during pregnancy may be a factor that leads to high percentage of population who are not US citizens.	2	16	2
16. High percentage of adults age 65+ with all teeth extracted may be a factor that leads to high percentage of adults with hypertension.	2	16	2
17. High percentage of men who visited a dentist regularly may be a factor that leads to low number of deaths due to diabetes per 100,000 population.	2	16	2
18. Low motor vehicle crash deaths may be a factor that leads to low rate of teenage pregnancy.	2	16	2
19. High percentage of adults with hypertension may be a factor that leads to low percentage of households with internet access.	4	15	1
20. High motor vehicle crash deaths may be a factor that leads to low access to exercise opportunities.	4	15	1

21. Low percentage of men who visited a dentist regularly may be a factor that leads to high number of people who are in jail (per 100,000 residents).	3	15	2
22. High access to exercise opportunities may be a factor that leads to high percent of households with computer and smartphone.	5	14	1
23. High percentage of people who go to work by walking may be a factor that leads to high access to dental care.	4	14	2
24. Low percentage of households with internet access may be a factor that leads to high percentage of adults who smoke.	3	14	3
25. Low rate of women who have cervical cancer may be a factor that leads to low percentage of children in poverty.	5	13	2
26. Low percentage of men who visited a dentist regularly may be a factor that leads to low median household income.	5	13	2
27. High percentage of adults age 65+ with all teeth extracted may be a factor that leads to high percentage of people with disabilities.	5	12	3
28. High percentage of adults with hypertension may be a factor that leads to low employment rate.	4	12	4
29. Low percentage of infants who were breastfed may be a factor that leads to high percentage of children in poverty.	5	11	4
30. High percentage of public-school students who use marijuana may be a factor that leads to low percentage of adults with children.	5	11	4

B.3 Ranked List of 30 Hard-To-Tell Claims (In Descending Order of Entropy)

	R	U	N
1. High percentage of adults who smoke may be a factor that leads to high percentage of people with vision difficulty.	8	5	7
2. Low percentage of infants who were breastfed may be a factor that leads to low average life expectancy.	7	8	5
3. Low percentage of infants who were breastfed may be a factor that leads to high percentage of adults with diabetes.	7	5	8

4. High employment rate may be a factor that leads to low percentage of adults with poor physical health.	8	7	5
5. High percentage of adults who smoke may be a factor that leads to high infant mortality.	9	6	5
6. Low percentage of infants who were breastfed may be a factor that leads to high percentage of adults with hypertension.	5	9	6
7. Low percentage of people who lack adequate access to food may be a factor that leads to low rate of deaths due to cervical cancer.	6	9	5
8. Low percentage of adults who exercise regularly may be a factor that leads to high percentage of preterm births.	8	8	4
9. High percentage of white people who have low income may be a factor that leads to high percentage of adults with poor mental health.	9	7	4
10. Low percentage of households with internet access may be a factor that leads to low average life expectancy.	5	10	5
11. High percentage of adults with at least college education may be a factor that leads to low percentage of adults with hypertension.	6	10	4
12. High rate of teenage pregnancy may be a factor that leads to high percentage of people who lack adequate access to food.	10	7	3
13. High percentage of people who are extremely poor may be a factor that leads to high percentage of preterm births.	10	7	3
14. High rate of teenage pregnancy may be a factor that leads to high child mortality.	11	4	5
15. Low median household income may be a factor that leads to low average life expectancy.	11	5	4
16. High homicide rate may be a factor that leads to high child mortality.	10	8	2
17. High percentage of children in single-parent households may be a factor that leads to high percentage of live births with low birthweight.	3	12	5
18. High percentage of people with hearing difficulty may be a factor that leads to high number of deaths due to injury.	12	5	3
19. High percentage of adults with hypertension may be a factor that leads to high rate of deaths due to cervical cancer.	5	12	3

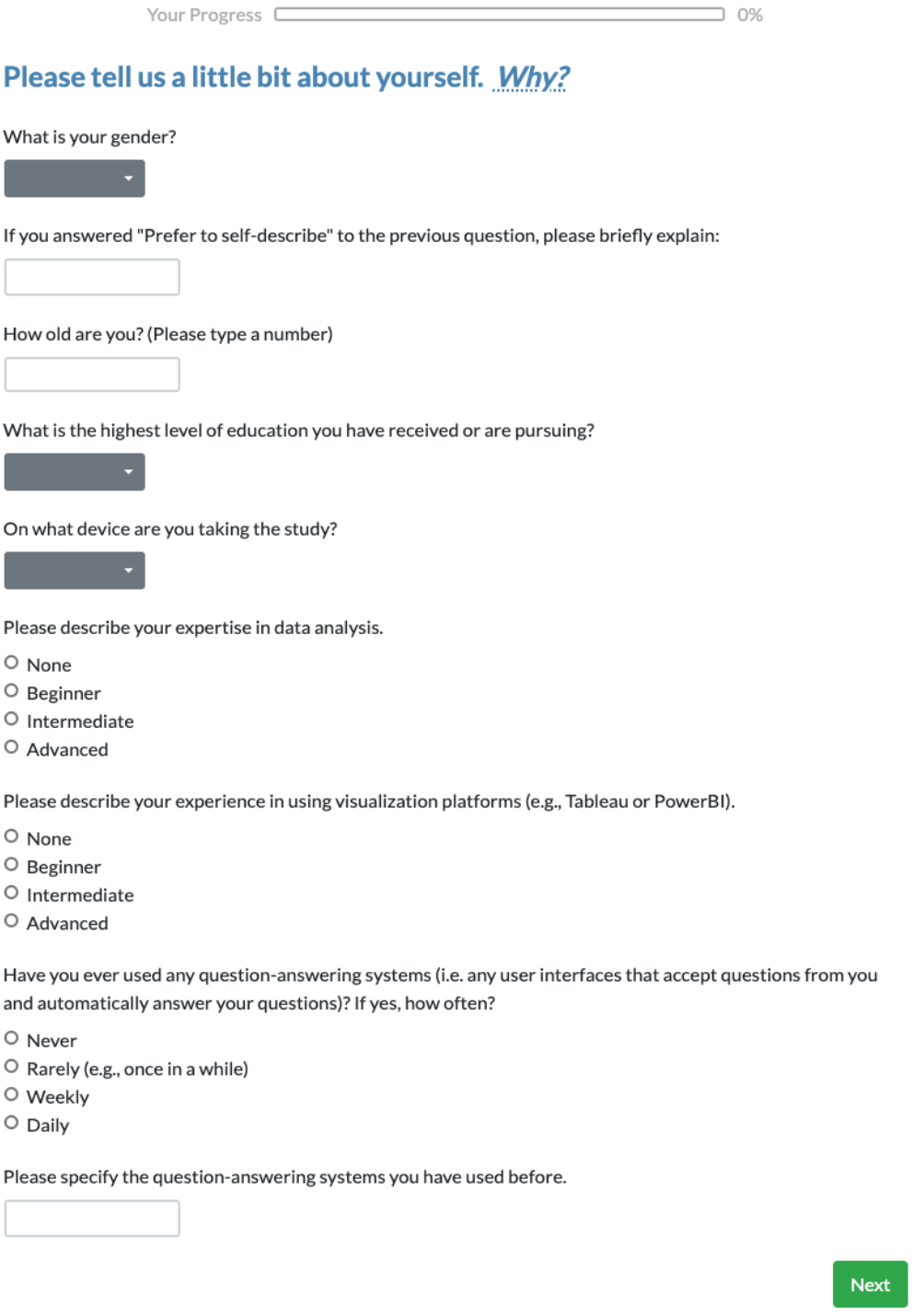
20. High percentage of people born in a foreign country may be a factor that leads to high median rent.	4	13	3
21. High percentage of live births with low birthweight may be a factor that leads to high homicide rate.	3	13	4
22. Low rate of teenage pregnancy may be a factor that leads to high average ACT Math score.	5	13	2
23. Low percentage of adults with at least college education may be a factor that leads to low percentage of men who visited a dentist regularly.	4	14	2
24. High percentage of adults with at least college education may be a factor that leads to low rate of deaths due to cervical cancer.	1	13	6
25. High percentage of men who visited a dentist regularly may be a factor that leads to low death rate per 100,000 population.	6	13	1
26. High average commute to work may be a factor that leads to low suicide rate.	1	14	5
27. High percentage of adults who exercise regularly may be a factor that leads to high percentage of people who work at home.	2	15	3
28. High access to exercise opportunities may be a factor that leads to high median rent.	3	15	2
29. High percentage of people born in a foreign country may be a factor that leads to low percentage of women who smoke during pregnancy.	2	16	2
30. Low access to exercise opportunities may be a factor that leads to low percentage of people born in a foreign country.	3	17	0

APPENDIX C

SCREENSHOTS OF INTERFACE USED IN THE STUDIES

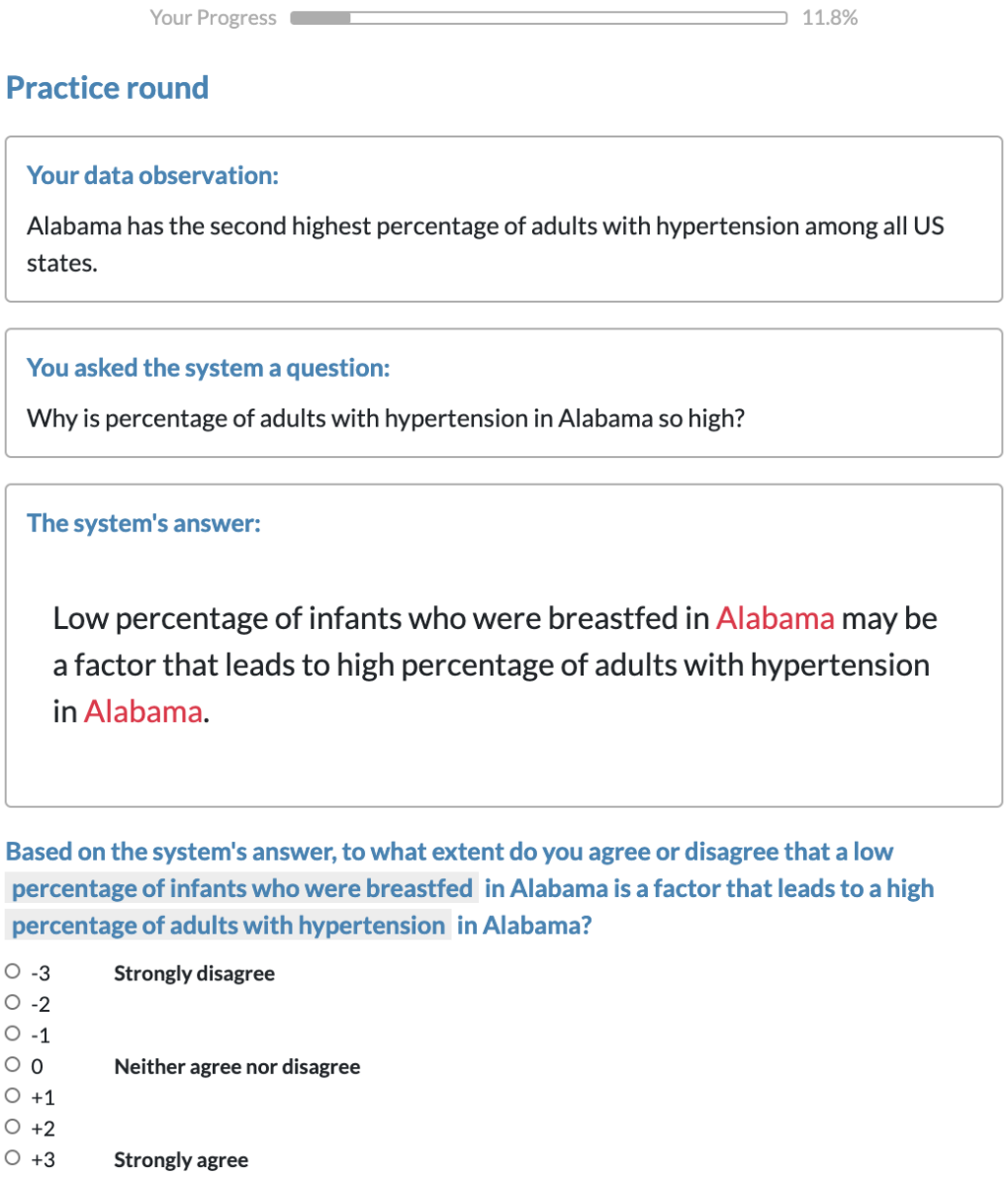
In Chapter 5, the interfaces for study 1 and study 2 were pretty much same. The only difference between the studies was the plausibility of causal claims. Across experimental conditions (i.e., claim only, claim + vis, claim + vis + description, claim + vis + description + warning), the designs of answers to why questions were different. To support replication, I describe the interface for the experiments in the following pages.

In both studies, regardless of experimental condition, participants first filled out a demographic survey.

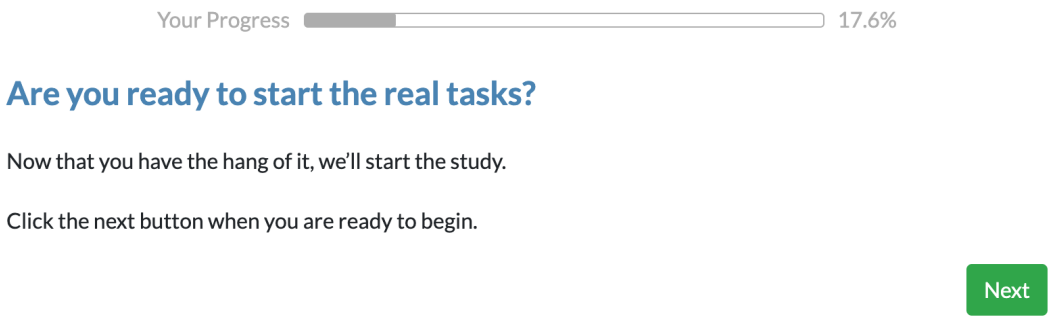


They then read the task instructions. Depending on the experimental condition, the instruction contents were different. The following shows the instructions for the claim-only condition in study 1.

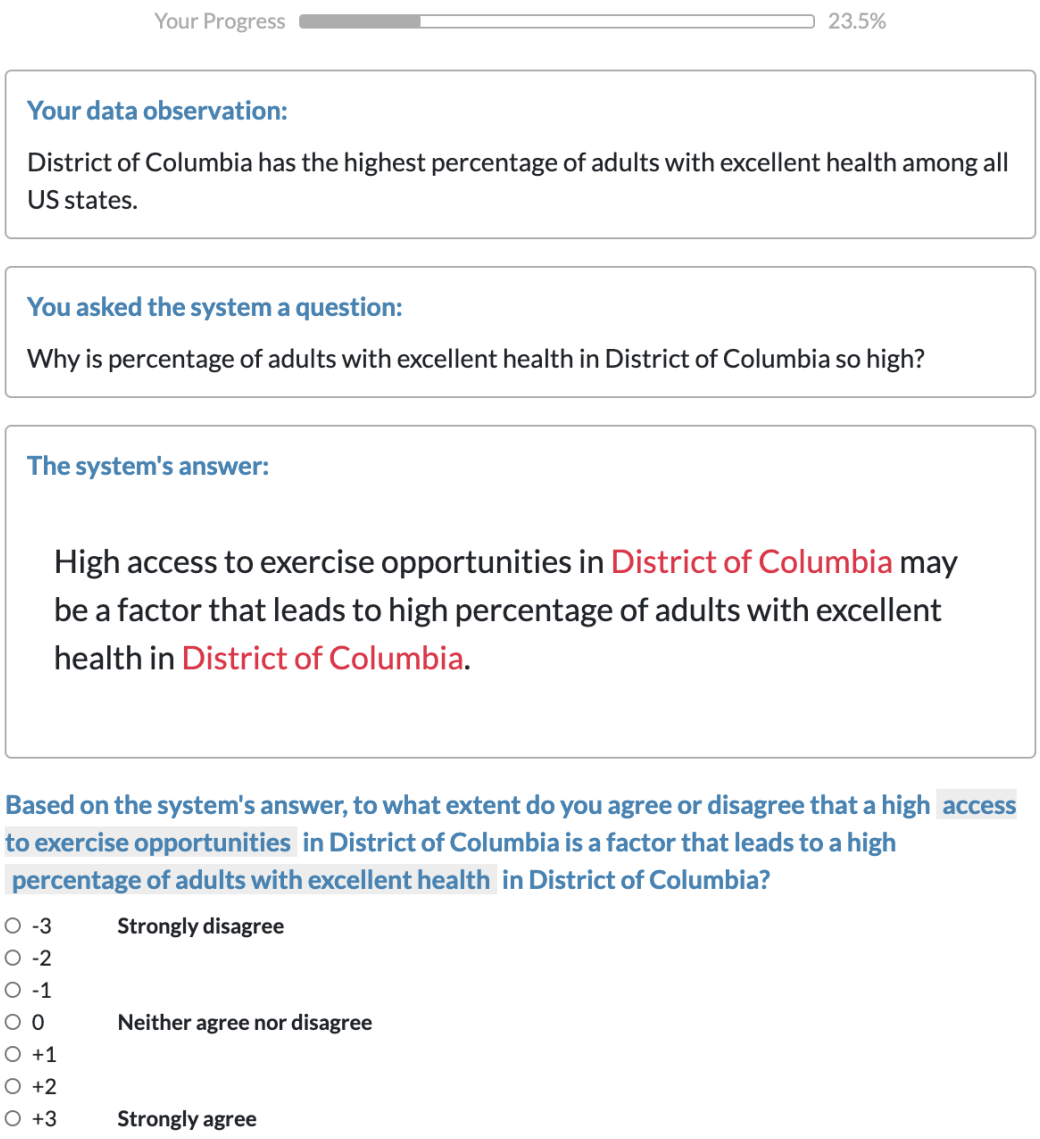
Following the instructions, they did a practice task. Below is the practice task for the claim-only condition in study 1.



When participants were ready for the actual tasks, they clicked on the next button.



Next, they reviewed nine answers to why questions. The designs of the answers were different across the four conditions. Here is the claim-only condition.



Here is the claim + vis condition.

Your Progress 23.5%

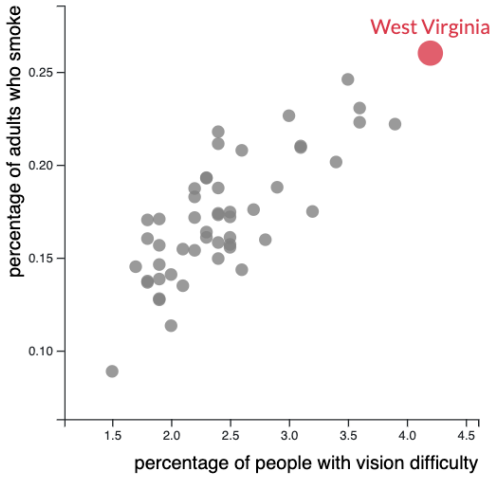
Your data observation:

West Virginia has the highest percentage of people with vision difficulty among all US states.

You asked the system a question:

Why is percentage of people with vision difficulty in West Virginia so high?

The system's answer (each dot in the scatterplot is a US state):



High percentage of adults who smoke in West Virginia may be a factor that leads to high percentage of people with vision difficulty in West Virginia.

Based on the system's answer, to what extent do you agree or disagree that a high percentage of adults who smoke in West Virginia is a factor that leads to a high percentage of people with vision difficulty in West Virginia?

- ○ -3 **Strongly disagree**
- ○ -2
- ○ -1
- ○ 0 **Neither agree nor disagree**
- ○ +1
- ○ +2
- ○ +3 **Strongly agree**

Here is the claim + vis + description condition.

Your Progress 23.5%

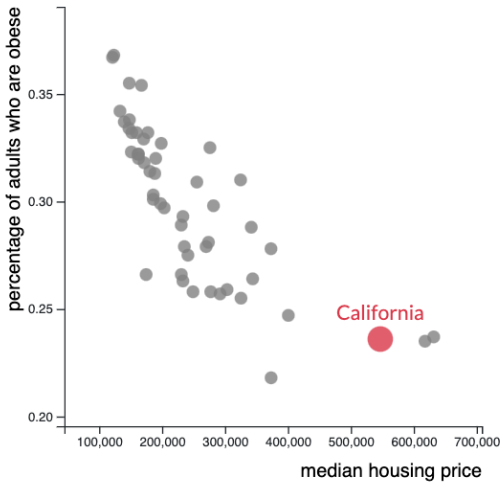
Your data observation:

California has the third highest median housing price among all US states.

You asked the system a question:

Why is median housing price in California so high?

The system's answer (each dot in the scatterplot is a US state):



As percentage of adults who are obese decreases, median housing price tends to increase.

Low percentage of adults who are obese in California may be a factor that leads to high median housing price in California.

Based on the system's answer, to what extent do you agree or disagree that a low percentage of adults who are obese in California is a factor that leads to a high median housing price in California?

- ○ -3 **Strongly disagree**
- ○ -2
- ○ -1
- ○ 0 **Neither agree nor disagree**
- ○ +1
- ○ +2
- ○ +3 **Strongly agree**

Here is the claim + vis + description + warning condition.

Your Progress 23.5%

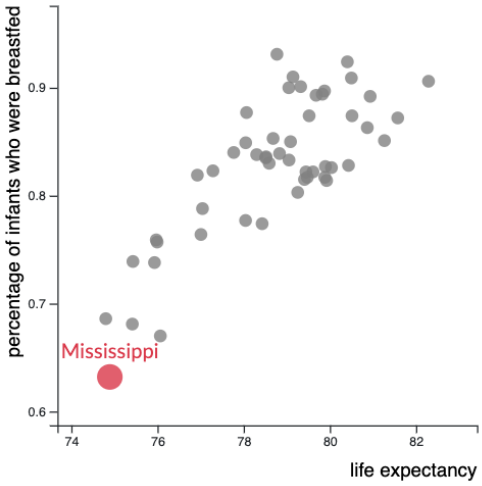
Your data observation:

Mississippi has the second lowest life expectancy among all US states.

You asked the system a question:

Why is life expectancy in Mississippi so low?

The system's answer (each dot in the scatterplot is a US state):



As percentage of infants who were breastfed decreases, life expectancy tends to decrease.

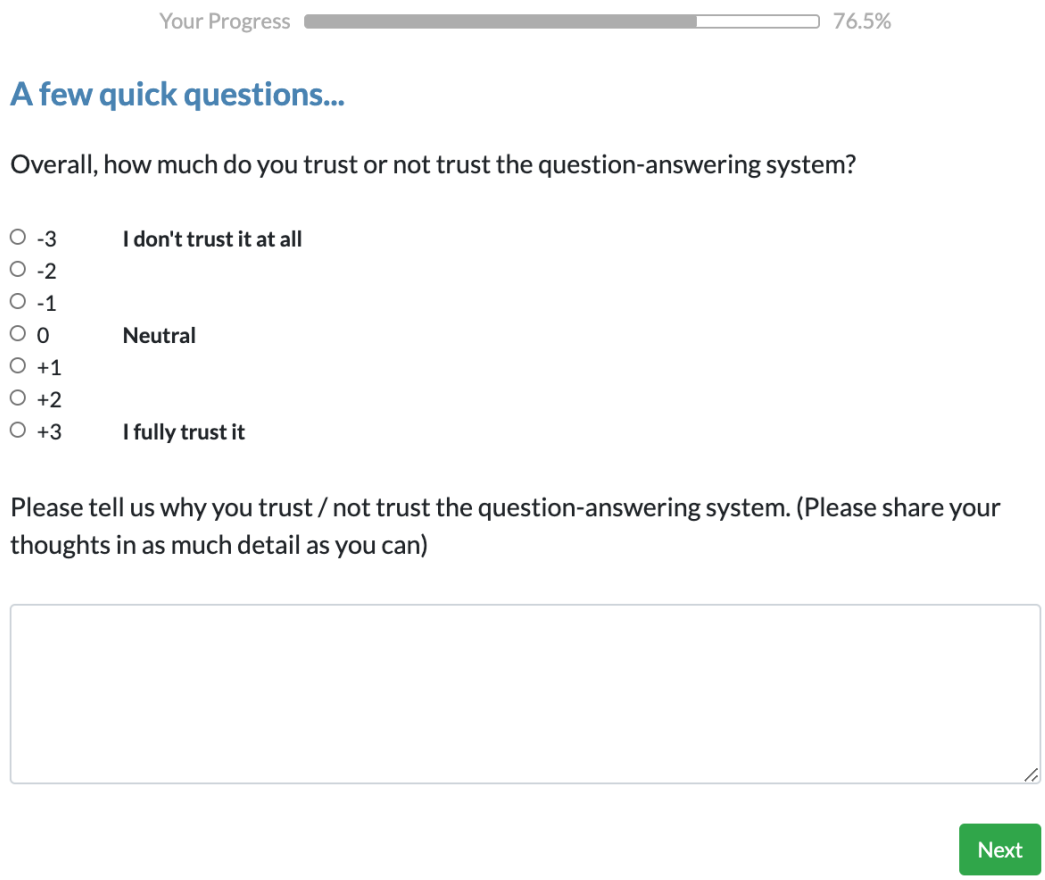
Low percentage of infants who were breastfed in Mississippi may be a factor that leads to low life expectancy in Mississippi.

The scatterplot only shows a correlation between percentage of infants who were breastfed and life expectancy. Correlation, however, does not imply causation. Interpret my answer with care!

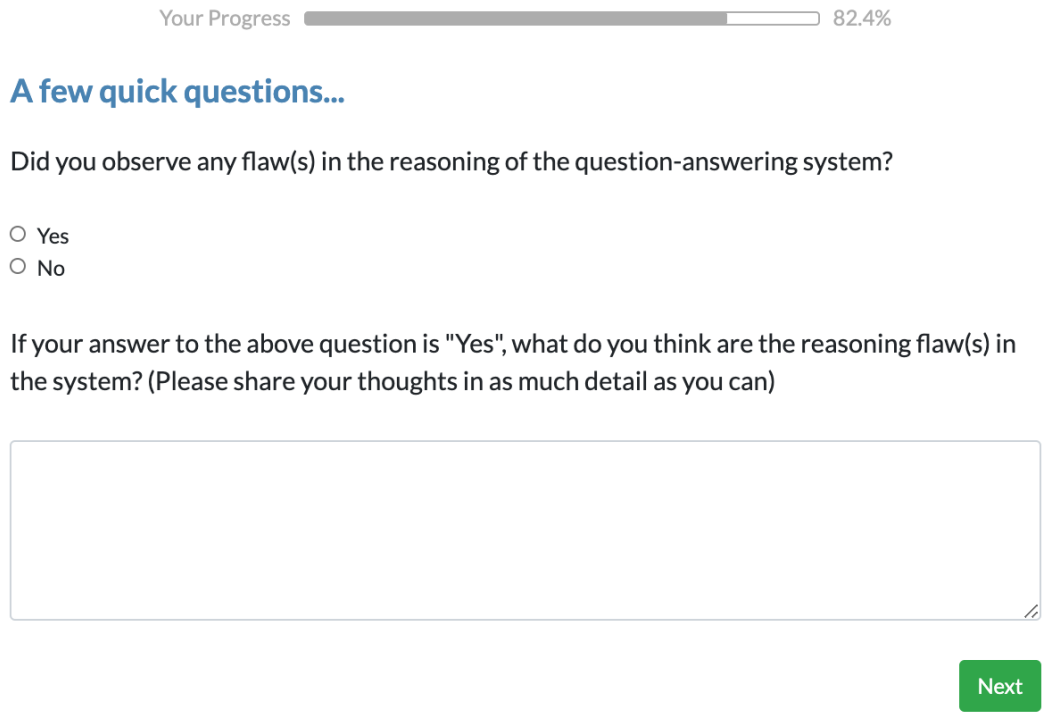
Based on the system's answer, to what extent do you agree or disagree that a low percentage of infants who were breastfed in Mississippi is a factor that leads to a low life expectancy in Mississippi?

- ○ -3 **Strongly disagree**
- ○ -2
- ○ -1
- ○ 0 **Neither agree nor disagree**
- ○ +1
- ○ +2
- ○ +3 **Strongly agree**

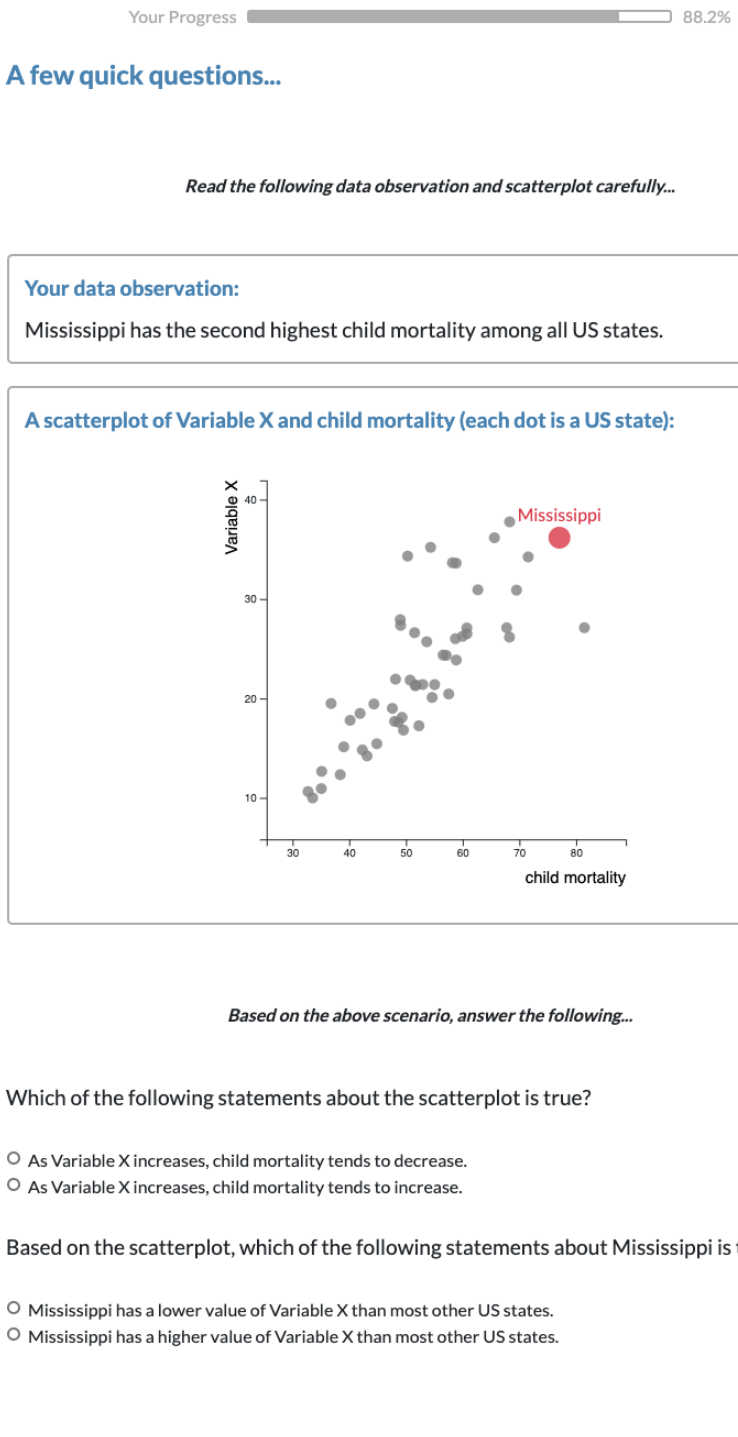
After the tasks, they rated their trust in the system and explained why they trusted or not trusted the system.



Participants described whether they identified any reasoning flaws in the system.



They then saw a data observation along with a scatterplot. This page has some questions for assessing participants' visualization literacy. Participants who failed to answer any of the questions were removed from subsequent analysis.



On the final page, I asked participants whether the scatterplot implied a causal relationship between variables. Participants further explained their answer.

Your Progress 94.1%

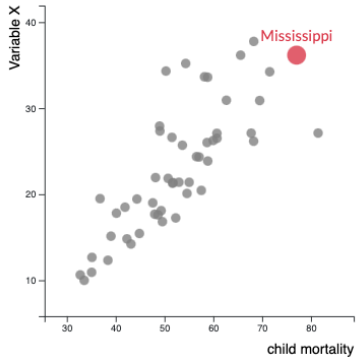
A few quick questions...

Read the following data observation and scatterplot carefully...

Your data observation:

Mississippi has the second highest child mortality among all US states.

A scatterplot of Variable X and child mortality (each dot is a US state):



Based on the above scenario, answer the following...

To what extent do you agree or disagree with this statement: **"THE SCATTERPLOT IMPLIES that the high value of Variable X is a factor that leads to the high child mortality in Mississippi."**

- ○ -3 **Strongly disagree**
- ○ -2
- ○ -1
- ○ 0 **Neither agree nor disagree**
- ○ +1
- ○ +2
- ○ +3 **Strongly agree**

Why do you agree or disagree with the statement? (Please share your thoughts in as much detail as you can)

Next

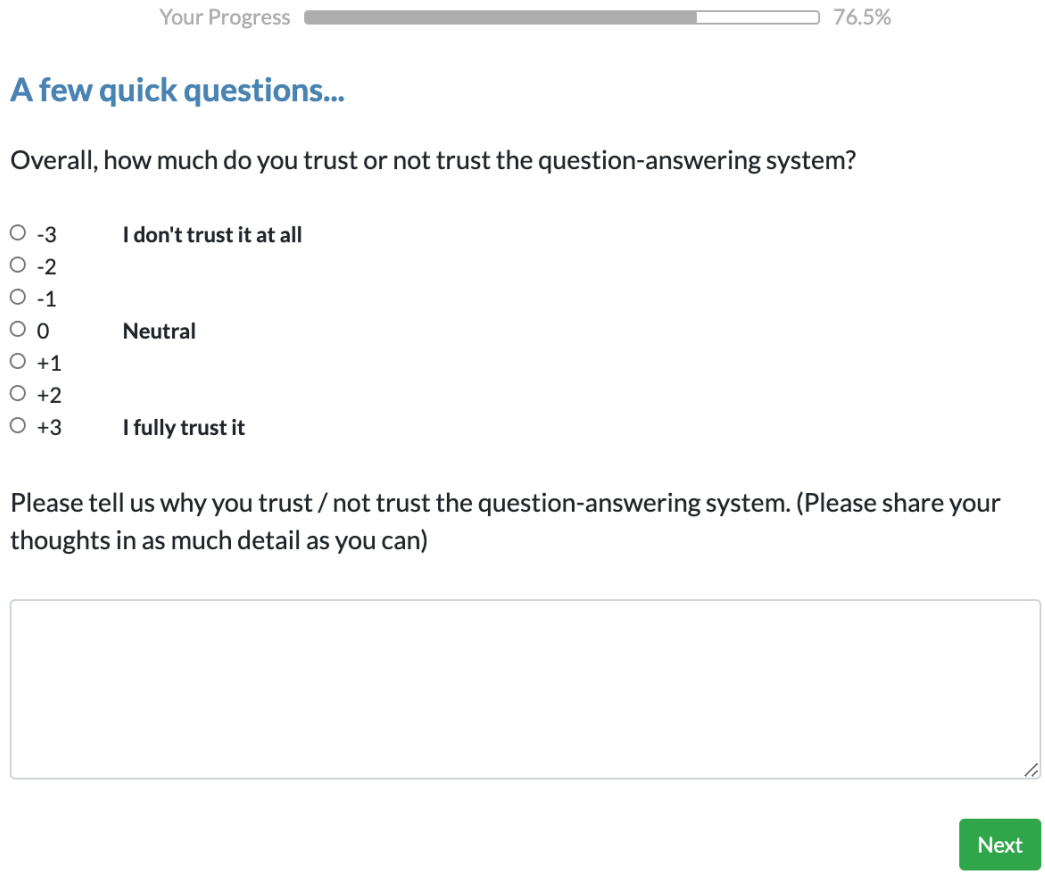
APPENDIX D

DETAILED QUALITATIVE CODING RESULTS

In Chapter 5, both study 1 and study 2 had three open-ended questions: Why trust or not trust the question-answering system? What are the reasoning flaws in the system? Why agree or disagree that a scatterplot implies causation? I present the detailed qualitative coding results in the following pages.

D.1 Study 1: Why Trust or Not Trust the System?

During study 1, participants were asked the following question:



The following pages summarize the categories of open-ended responses across the four experimental conditions. I conducted Fisher's exact tests to determine if the categories were significantly different across conditions.

Some answers do not make sense

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	28	22	50
Claim + vis	20	30	50
Claim + vis + description	17	33	50
Claim + vis + description + warning	16	34	50

The number of responses that contained "some answers do not make sense" did not significantly differ across conditions (p=.064).

Some answers make sense

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	16	34	50
Claim + vis	12	38	50
Claim + vis + description	12	38	50
Claim + vis + description + warning	12	38	50

The number of responses that contained "some answers make sense" did not significantly differ across conditions (p=.76).

The system shows the data

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	0	50	50
Claim + vis	5	45	50
Claim + vis + description	7	43	50
Claim + vis + description + warning	7	43	50

The number of responses that contain "the system shows the data" significantly differed across conditions (p=.020). I conducted six post-hoc Fisher's exact tests with Holm–Bonferroni correction but did not observe a significant difference between any of the pairs.

The system confuses correlation and causation

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	1	49	50
Claim + vis	5	45	50
Claim + vis + description	8	42	50
Claim + vis + description + warning	4	46	50

The number of responses that contained "the system confuses correlation and causation" did not significantly differ across conditions (p=.096).

The system does not provide enough support for its answers

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	2	48	50
Claim + vis	5	45	50
Claim + vis + description	3	47	50
Claim + vis + description + warning	5	45	50

The number of responses that contained "the system does not provide enough support for its answers" did not significantly differ across conditions (p=.64).

The system provides support for its answers

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	1	49	50
Claim + vis	4	46	50
Claim + vis + description	4	46	50
Claim + vis + description + warning	4	46	50

The number of responses that contained "the system provides support for its answers" did not significantly differ across conditions (p=.51).

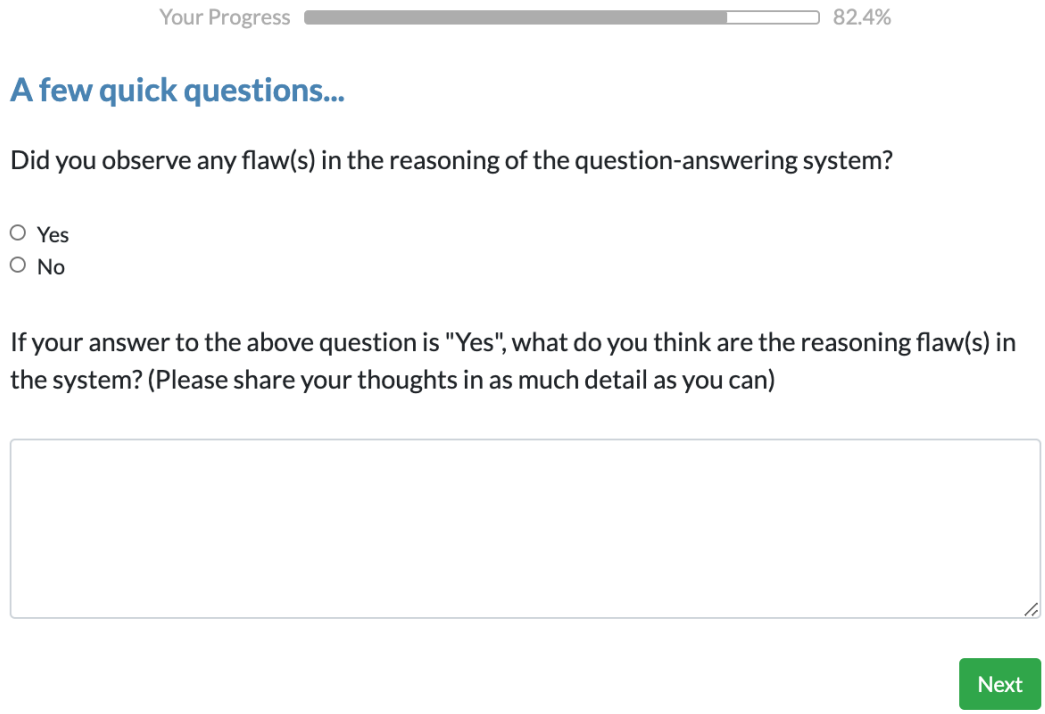
The system shows a warning

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	0	50	50
Claim + vis	0	50	50
Claim + vis + description	0	50	50
Claim + vis + description + warning	2	48	50

The number of responses that contained "the system shows a warning" did not significantly differ across conditions (p=.25).

D.2 Study 1: What Are the Reasoning Flaws in the System?

During study 1, participants were asked the following question:



The following pages summarize the categories of open-ended responses across the four experimental conditions. I conducted Fisher's exact tests to determine if the categories were significantly different across conditions.

Some answers do not make sense

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants who observed reasoning flaws
Claim only	24	4	28
Claim + vis	18	7	25
Claim + vis + description	15	10	25
Claim + vis + description + warning	11	8	19

The number of responses that contained "some answers do not make sense" did not significantly differ across conditions (p=.10).

The system confuses correlation and causation

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants who observed reasoning flaws
Claim only	2	26	28
Claim + vis	5	20	25
Claim + vis + description	5	20	25
Claim + vis + description + warning	3	16	19

The number of responses that contained "the system confuses correlation and causation" did not significantly differ across conditions (p=.49).

The system does not provide enough support for its answers

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants who observed reasoning flaws
Claim only	1	27	28
Claim + vis	0	25	25
Claim + vis + description	4	21	25
Claim + vis + description + warning	3	16	19

The number of responses that contained "the system does not provide enough support for its answers" did not significantly differ across conditions (p=.065).

The system considers only one factor

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants who observed reasoning flaws
Claim only	0	28	28
Claim + vis	0	25	25
Claim + vis + description	1	24	25
Claim + vis + description + warning	2	17	19

The number of responses that contained "the system considers only one factor" did not significantly differ across conditions (p=.12).

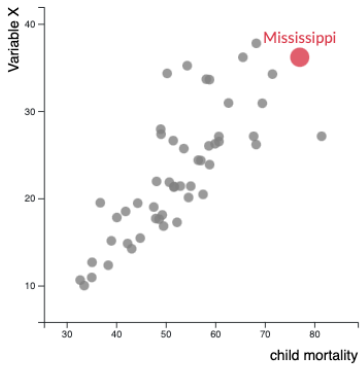
D.3 Study 1: Why Agree or Disagree That a Scatterplot Implies Causation?

During study 1, participants were asked the following questions:

Your data observation:

Mississippi has the second highest child mortality among all US states.

A scatterplot of Variable X and child mortality (each dot is a US state):



Based on the above scenario, answer the following...

To what extent do you agree or disagree with this statement: **"THE SCATTERPLOT IMPLIES that the high value of Variable X is a factor that leads to the high child mortality in Mississippi."**

- ○ -3 **Strongly disagree**
- ○ -2
- ○ -1
- ○ 0 **Neither agree nor disagree**
- ○ +1
- ○ +2
- ○ +3 **Strongly agree**

Why do you agree or disagree with the statement? (Please share your thoughts in as much detail as you can)

The following pages summarize the categories of open-ended responses across the four experimental conditions. I conducted Fisher's exact tests to determine if the categories were significantly different across conditions.

The scatterplot shows a correlation

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	28	22	50
Claim + vis	22	28	50
Claim + vis + description	25	25	50
Claim + vis + description + warning	17	33	50

The number of responses that contained "the scatterplot shows a correlation" did not significantly differ across conditions (p=.15).

Correlation is not causation

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	5	45	50
Claim + vis	4	46	50
Claim + vis + description	2	48	50
Claim + vis + description + warning	6	44	50

The number of responses that contained "correlation is not causation" did not significantly differ across conditions (p=.59).

Variable X is unknown, and I cannot judge

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	1	49	50
Claim + vis	4	46	50
Claim + vis + description	5	45	50
Claim + vis + description + warning	7	43	50

The number of responses that contained "variable X is unknown, and I cannot judge" did not significantly differ across conditions (p=.17).

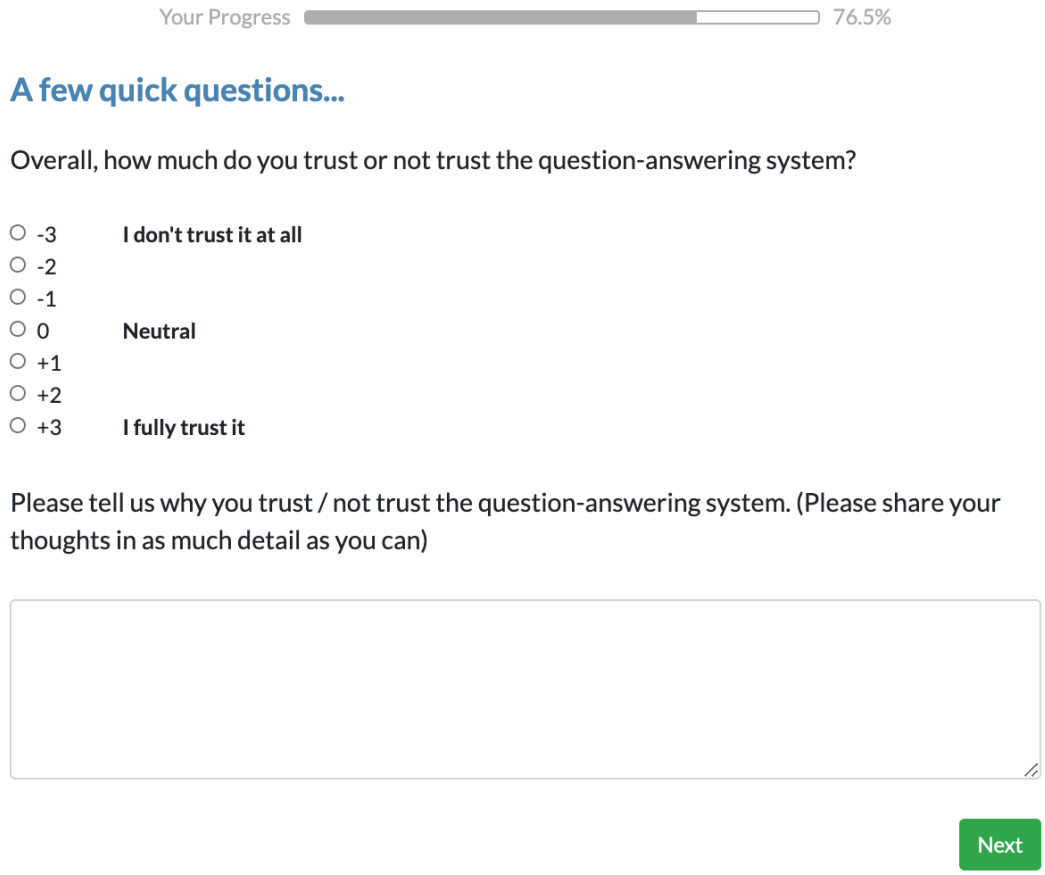
There are outliers in the scatterplot

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	2	48	50
Claim + vis	2	48	50
Claim + vis + description	1	49	50
Claim + vis + description + warning	3	47	50

The number of responses that contained "there are outliers in the scatterplot" did not significantly differ across conditions (p=.96).

D.4 Study 2: Why Trust or Not Trust the System?

During study 2, participants were asked the following question:



The following pages summarize the categories of open-ended responses across the four experimental conditions. I conducted Fisher's exact tests to determine if the categories were significantly different across conditions.

The answers make sense

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	28	22	50
Claim + vis	20	30	50
Claim + vis + description	13	37	50
Claim + vis + description + warning	16	34	50

The number of responses that contained "the answers make sense" significantly differed across conditions (p=.014). I conducted six post-hoc comparisons using Fisher's exact tests with Holm–Bonferroni correction. I only observed a significant difference between claim only and claim + vis + description (p=.025).

The system provides support for its answers

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	4	46	50
Claim + vis	9	41	50
Claim + vis + description	13	37	50
Claim + vis + description + warning	13	37	50

The number of responses that contained "the system provides support for its answers" did not significantly differ across conditions (p=.060).

The system does not provide enough support for its answers

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	6	44	50
Claim + vis	3	47	50
Claim + vis + description	4	46	50
Claim + vis + description + warning	3	47	50

The number of responses that contained "the system does not provide enough support for its answers" did not significantly differ across conditions (p=.77).

The system considers only one factor

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	2	48	50
Claim + vis	5	45	50
Claim + vis + description	2	48	50
Claim + vis + description + warning	5	45	50

The number of responses that contained "the system considers only one factor" did not significantly differ across conditions (p=.48).

The system shows the data

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	0	50	50
Claim + vis	6	44	50
Claim + vis + description	6	44	50
Claim + vis + description + warning	4	46	50

The number of responses that contained "the system shows the data" significantly differed across conditions (p=.042). I conducted six post-hoc comparisons using Fisher's exact tests with Holm–Bonferroni correction and did not observe a significant difference between any pairs.

The system confuses correlation and causation

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	2	48	50
Claim + vis	3	47	50
Claim + vis + description	0	50	50
Claim + vis + description + warning	0	50	50

The number of responses that contained "the system confuses correlation and causation" did not significantly differ across conditions (p=.17).

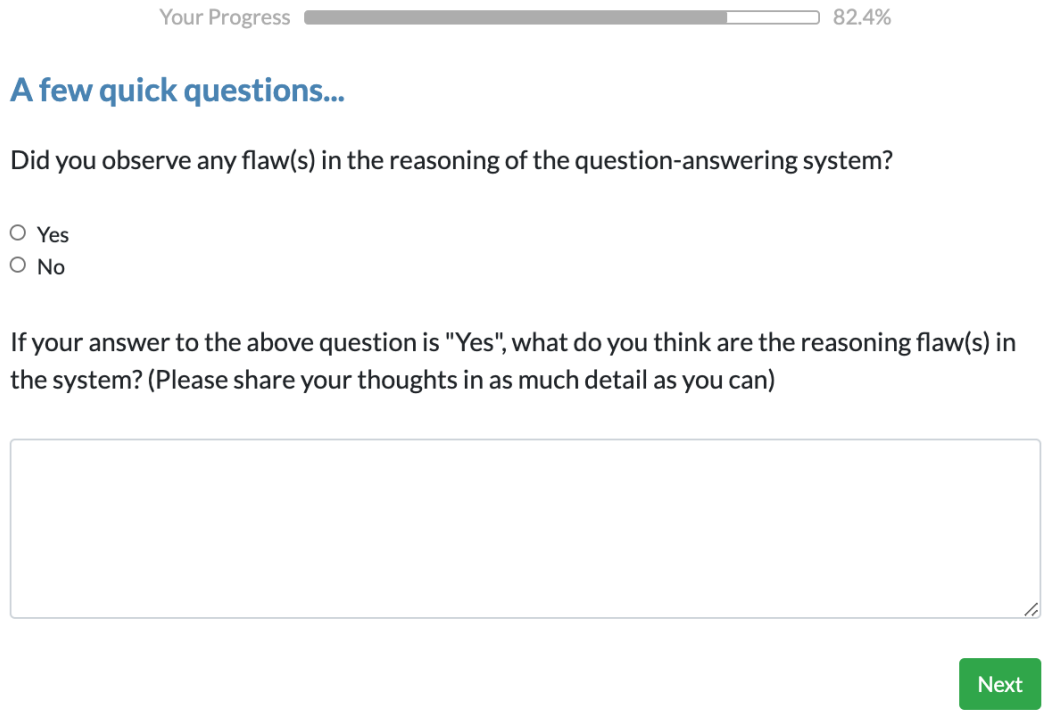
The system shows a warning

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	0	50	50
Claim + vis	0	50	50
Claim + vis + description	0	50	50
Claim + vis + description + warning	3	47	50

The number of responses that contained "the system shows a warning" did not significantly differ across conditions (p=.060).

D.5 Study 2: What Are the Reasoning Flaws in the System?

During study 2, participants were asked the following question:



The following pages summarize the categories of open-ended responses across the four experimental conditions. I conducted Fisher's exact tests to determine if the categories were significantly different across conditions.

The system considers only one factor

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants who observed reasoning flaws
Claim only	1	6	7
Claim + vis	6	8	14
Claim + vis + description	1	6	7
Claim + vis + description + warning	2	5	7

The number of responses that contained "the system considers only one factor" did not significantly differ across conditions (p=.48).

The system confuses correlation and causation

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants who observed reasoning flaws
Claim only	2	5	7
Claim + vis	3	11	14
Claim + vis + description	2	5	7
Claim + vis + description + warning	0	7	7

The number of responses that contained "the system confuses correlation and causation" did not significantly differ across conditions (p=.66).

The system does not provide enough support for its answers

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants who observed reasoning flaws
Claim only	0	7	7
Claim + vis	2	12	14
Claim + vis + description	1	6	7
Claim + vis + description + warning	3	4	7

The number of responses that contained "the system does not provide enough support for its answers" did not significantly differ across conditions (p=.26).

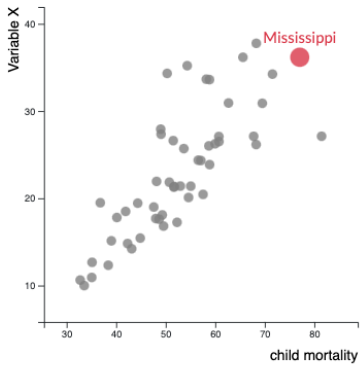
D.6 Study 2: Why Agree or Disagree That a Scatterplot Implies Causation?

During study 2, participants were asked the following questions:

Your data observation:

Mississippi has the second highest child mortality among all US states.

A scatterplot of Variable X and child mortality (each dot is a US state):



Based on the above scenario, answer the following...

To what extent do you agree or disagree with this statement: **"THE SCATTERPLOT IMPLIES that the high value of Variable X is a factor that leads to the high child mortality in Mississippi."**

- ○ -3 **Strongly disagree**
- ○ -2
- ○ -1
- ○ 0 **Neither agree nor disagree**
- ○ +1
- ○ +2
- ○ +3 **Strongly agree**

Why do you agree or disagree with the statement? (Please share your thoughts in as much detail as you can)

The following pages summarize the categories of open-ended responses across the four experimental conditions. I conducted Fisher's exact tests to determine if the categories were significantly different across conditions.

The scatterplot shows a correlation

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	18	32	50
Claim + vis	24	26	50
Claim + vis + description	18	32	50
Claim + vis + description + warning	27	23	50

The number of responses that contained "the scatterplot shows a correlation" did not significantly differ across conditions (p=.18).

Correlation is not causation

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	4	46	50
Claim + vis	3	47	50
Claim + vis + description	1	49	50
Claim + vis + description + warning	7	43	50

The number of responses that contained "correlation is not causation" did not significantly differ across conditions (p=.16).

There are outliers in the scatterplot

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	2	48	50
Claim + vis	3	47	50
Claim + vis + description	4	46	50
Claim + vis + description + warning	5	45	50

The number of responses that contained "there are outliers in the scatterplot" did not significantly differ across conditions (p=.79).

Variable X is unknown, and I cannot judge

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	3	47	50
Claim + vis	2	48	50
Claim + vis + description	2	48	50
Claim + vis + description + warning	4	46	50

The number of responses that contained "variable X is unknown, and I cannot judge" did not significantly differ across conditions (p=.90).

The dots in the scatterplot are disperse

Condition	Number of responses that contain the category	Number of responses that do not contain the category	Number of participants per condition
Claim only	0	50	50
Claim + vis	0	50	50
Claim + vis + description	2	48	50
Claim + vis + description + warning	1	49	50

The number of responses that contained “the dots in the scatterplot are disperse” did not significantly differ across conditions (p=.62).

REFERENCES

[1] R. L. Ackoff, "From data to wisdom," *Journal of Applied Systems Analysis*, vol. 16, pp. 3–9, 1989.

[2] P. D. Adamczyk and B. P. Bailey, "If not now, when? the effects of interruption at different moments within task execution," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2004, pp. 271–278.

[3] S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst, "Futzing and moseying: Interviews with professional data analysts on exploration practices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 22–31, 2018.

[4] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," in *IEEE Symposium on Information Visualization*, IEEE, 2005, pp. 111–117.

[5] A. Anand and J. Talbot, "Automatic selection of partitioning variables for small multiple displays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 669–677, 2015.

[6] Automated Insights. (2020). "Automated Insights: Natural Language Generation." `https://automatedinsights.com`.

[7] M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck, "Feedback-driven interactive exploration of large multidimensional data supported by visual classifier," in *IEEE Conference on Visual Analytics Science and Technology*, IEEE, 2014, pp. 43–52.

[8] C. Bergland. (2019). "Where Our "Eureka!" Moments Come From | Psychology Today." `https://www.psychologytoday.com/us/blog/...`

[9] E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery," in *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, 2009, pp. 12–20.

[10] W. M. Bolstad and J. M. Curran, *Introduction to Bayesian Statistics*. John Wiley & Sons, 2016.

[11] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 519–528, 2015.

[12] J. Boy, A. V. Pandey, J. Emerson, M. Satterthwaite, O. Nov, and E. Bertini, "Showing people behind data: Does anthropomorphizing visualizations elicit more empathy for human rights data?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2017, pp. 5462–5474.

[13] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang, "Finding waldo: Learning about users from their interactions," *IEEE Transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1663–1672, 2014.

[14] C. Bryan, K.-L. Ma, and J. Woodring, "Temporal Summary Images: An approach to narrative visualization via interactive annotation generation and placement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 511–520, 2016.

[15] M. J. Cafarella, A. Halevy, and J. Madhavan, "Structured data on the web," *Communications of the ACM*, vol. 54, no. 2, pp. 72–79, 2011.

[16] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, "Webtables: Exploring the power of tables on the web," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 538–549, 2008.

[17] Cambridge University Press. (2020). "INSIGHT | definition in the Cambridge English Dictionary."
`https://dictionary.cambridge.org/us/dictionary/...`

[18] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

[19] W. Carpenter, "The aha! moment: The science behind creative insights," in *Toward Super-Creativity-Improving Creativity in Humans, Machines, and Human-Machine Collaborations*, IntechOpen, 2019.

[20] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski, "Characterizing guidance in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 111–120, 2016.

[21] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky, "Defining insight for visual analytics," *IEEE Computer Graphics and Applications*, vol. 29, no. 2, pp. 14–17, 2009.

[22] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver, "Data, information, and knowledge in visualization," *IEEE Computer Graphics and Applications*, vol. 29, no. 1, pp. 12–19, 2008.

[23] Y. Chen, J. Yang, and W. Ribarsky, "Toward effective insight management in visual analytics systems," in *2009 IEEE Pacific Visualization Symposium*, IEEE, 2009, pp. 49–56.

[24] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–12.

[25] E. K. Choe, B. Lee, *et al.*, "Characterizing visualization insights from quantified selfers' personal data presentations," *IEEE Computer Graphics and Applications*, vol. 35, no. 4, pp. 28–37, 2015.

[26] I. K. Choi, T. Childers, N. K. Raveendranath, S. Mishra, K. Harris, and K. Reda, "Concept-driven visual analytics: An exploratory study of model-and hypothesis-based reasoning with visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[27] W. S. Cleveland, P. Diaconis, and R. McGill, "Variables on scatterplots look more highly correlated when the scales are increased," *Science*, vol. 216, no. 4550, pp. 1138–1141, 1982.

[28] J. Corbin and A. Strauss, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, 2014.

[29] M. Correll, "Ethical dimensions of visualization research," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–13.

[30] M. Correll, E. Bertini, and S. Franconeri, "Truncating the y-axis: Threat or menace?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2020, pp. 1–12.

[31] M. Correll and J. Heer, "Black hat visualization," in *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe)*, 2017.

[32] Z. Cui, S. K. Badam, M. A. Yalçin, and N. Elmqvist, "DataSite: Proactive visual data exploration with computation of insight-based recommendations," *Information Visualization*, vol. 18, no. 2, pp. 251–267, 2019.

[33] N. Dagnall. (2018). "Why do people believe in superstitions? It's all a trick of the mind — Quartz." `https://qz.com/1319382/why-do-people...`

[34] T. N. Dang and L. Wilkinson, "ScagExplorer: Exploring scatterplots by their scagnostics," in *2014 IEEE Pacific Visualization Symposium*, IEEE, 2014, pp. 73–80.

[35] Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati, "Foresight: Recommending visual insights," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1937–1940, 2017.

[36] E. Dimara, G. Bailly, A. Bezerianos, and S. Franconeri, "Mitigating the attraction effect with visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 850–860, 2018.

[37] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 2, pp. 1413–1432, 2020.

[38] R. Ding, S. Han, Y. Xu, H. Zhang, and D. Zhang, "QuickInsights: Quick and automatic discovery of insights from multi-dimensional data," in *Proceedings of the 2019 International Conference on Management of Data*, 2019, pp. 317–332.

[39] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang, "Recovering reasoning processes from user interactions," *IEEE Computer Graphics and Applications*, vol. 29, no. 3, pp. 52–61, 2009.

[40] F. Du, C. Plaisant, N. Spring, and B. Shneiderman, "Finding similar people to guide life choices: Challenge, design, and evaluation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2017, pp. 5498–5544.

[41] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: An empirical study of the effectiveness of web browser phishing warnings," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2008, pp. 1065–1074.

[42] U. Ehsan, B. Harrison, L. Chan, and M. O. Riedl, "Rationalization: A neural machine translation approach to generating natural language explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 81–87.

[43] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: A technique for explainable ai and its effects on human perceptions," in *Proceedings of the International Conference on Intelligent User Interfaces*, 2019, pp. 263–274.

[44] M. Eslami, S. R. Krishna Kumaran, C. Sandvig, and K. Karahalios, "Communicating algorithmic process in online behavioral advertising," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[45] E. Fast, W. McGrath, P. Rajpurkar, and M. S. Bernstein, "Augur: Mining human behaviors from fiction to power interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2016, pp. 237–247.

[46] T. Gao, J. R. Hullman, E. Adar, B. Hecht, and N. Diakopoulos, "NewsViews: An automated pipeline for creating custom geovisualizations for news," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 3005–3014.

[47] S. R. Gomez, H. Guo, C. Ziemkiewicz, and D. H. Laidlaw, "An insight-and task-based methodology for evaluating spatiotemporal visual analytics," in *IEEE Conference on Visual Analytics Science and Technology*, IEEE, 2014, pp. 63–72.

[48] D. Gotz and M. X. Zhou, "Characterizing users' visual analytic activity for insight provenance," *Information Visualization*, vol. 8, no. 1, pp. 42–55, 2009.

[49] H. Griffey. (2018). "The lost art of concentration: being distracted in a digital world | Health wellbeing | The Guardian." `https://www.theguardian.com/lifeandstyle/2018/...`

[50] J. Grudin, "Computer-supported cooperative work: History and focus," *Computer*, vol. 27, no. 5, pp. 19–26, 1994.

[51] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw, "A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 51–60, 2015.

[52] P. J. Guo, S. Kandel, J. M. Hellerstein, and J. Heer, "Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2011, pp. 65–74.

[53] J. Heer, "Agency plus automation: Designing artificial intelligence into interactive systems," *Proceedings of the National Academy of Sciences*, vol. 116, no. 6, pp. 1844–1850, 2019.

[54] J. Heer and M. Agrawala, "Multi-scale banking to 45 degrees," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 701–708, 2006.

[55] J. Heer, F. B. Viégas, and M. Wattenberg, "Voyagers and voyeurs: Supporting asynchronous collaborative information visualization," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 1029–1038.

[56] E. Hellier, D. B. Wright, J. Edworthy, and S. Newstead, "On the stability of the arousal strength of warning signal words," *Applied Cognitive Psychology: The Offi-*

cial Journal of the Society for Applied Research in Memory and Cognition, vol. 14, no. 6, pp. 577–592, 2000.

[57] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, ACM, 2000, pp. 241–250.

[58] M. A. Hernán and R. J. M, *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC, 2020.

[59] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999, pp. 159–166.

[60] K. Hu, D. Orghian, and C. Hidalgo, "DIVE: A mixed-initiative system supporting integrated data exploration workflows," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, 2018, pp. 1–7.

[61] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: Automatic generation of annotated stock visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2707–2716.

[62] J. Hullman and N. Diakopoulos, "Visualization rhetoric: Framing effects in narrative visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2231–2240, 2011.

[63] S. T. Iqbal and B. P. Bailey, "Effects of intelligent notification management on users and their tasks," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 93–102.

[64] R. J. Jacob, "The use of eye movements in human-computer interaction techniques: What you look at is what you get," *ACM Transactions on Information Systems*, vol. 9, no. 2, pp. 152–169, 1991.

[65] Kaiser Family Foundation. (2020). "KFF - Health Policy Analysis, Polling and Journalism." `https://www.kff.org`.

[66] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, "Profiler: Integrated statistical analysis and visualization for data quality assessment," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 2012, pp. 547–554.

[67] E. Kandogan, A. Balakrishnan, E. M. Haber, and J. S. Pierce, "From data to insight: Work practices of analysts in the enterprise," *IEEE Computer Graphics and Applications*, vol. 34, no. 5, pp. 42–50, 2014.

[68] A. Key, B. Howe, D. Perry, and C. Aragon, "VizDeck: Self-organizing dashboards for visual analytics," in *Proceedings of the 2012 International Conference on Management of Data*, 2012, pp. 681–684.

[69] D. H. Kim, E. Hoque, and M. Agrawala, "Answering questions about charts and generating visual explanations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[70] Y.-S. Kim, K. Reinecke, and J. Hullman, "Data through others' eyes: The impact of visualizing others' expectations on visualization interpretation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 760–769, 2017.

[71] R. F. Kizilcec, "How much information? effects of transparency on trust in an algorithmic interface," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2016, pp. 2390–2395.

[72] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 1: Alternative perspectives," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 70–73, 2006.

[73] ——, "Making sense of sensemaking 2: A macrocognitive model," *IEEE Intelligent Systems*, vol. 21, no. 5, pp. 88–92, 2006.

[74] H.-K. Kong, Z. Liu, and K. Karahalios, "Frames and slants in titles of visualizations on controversial topics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2018, pp. 1–12.

[75] ——, "Trust and recall of information across varying degrees of title-visualization misalignment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–13.

[76] P.-M. Law, R. C. Basole, and Y. Wu, "Duet: Helping data analysis novices conduct pairwise comparisons by minimal specification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 427–437, 2018.

[77] P.-M. Law, S. Das, and R. C. Basole, "Comparing apples and oranges: Taxonomy and design of pairwise comparisons within tabular data," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[78] P.-M. Law, A. Endert, and J. Stasko, "Characterizing automated data insights," *arXiv preprint arXiv:2008.13060*, 2020.

[79] ——, "What are data insights to professional visualization users?" *arXiv preprint arXiv:2008.13057*, 2020.

[80] P.-M. Law, L. Y.-H. Lo, A. Endert, J. Stasko, and H. Qu, "Causal perception in question-answering systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2021.

[81] D. J.-L. Lee. (2020). "Insight Machines: The Past, Present, and Future of Visualization Recommendation."
`https://medium.com/multiple-views-visualization...`

[82] D. J.-L. Lee, H. Dev, H. Hu, H. Elmeleegy, and A. Parameswaran, "Avoiding drill-down fallacies with VisPilot: Assisted exploration of data subsets," in *Proceedings of the International Conference on Intelligent User Interfaces*, 2019, pp. 186–196.

[83] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[84] E. L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*. New York, NY, USA: Springer-Verlag, 2006, ISBN: 978-0-387-35212-1.

[85] S. Lewandowsky, J. Cook, U. Ecker, A. Albarracín, M. A. Amazeen, P. Kendeou, D. Lombardi, E. J. Newman, G. Pennycook, E. Porter, D. G. Rand, D. N. Rapp, J. Reifler, J. Roozenbeek, P. Schmid, C. M. Seifert, G. M. Sinatra, B. Swire-Thompson, S. van der Linden, E. K. Vraga, T. J. Wood, and M. S. Zaragoza, *The Debunking Handbook 2020*. 2020.

[86] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological Science in the Public Interest*, vol. 13, no. 3, pp. 106–131, 2012.

[87] ——, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological Science in the Public Interest*, vol. 13, no. 3, pp. 106–131, 2012.

[88] A. Liew, "DIKIW: Data, information, knowledge, intelligence, sisdom and their interrelationships," *Business Management Dynamics*, vol. 2, no. 10, p. 49, 2013.

[89] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 2119–2128.

[90] C. Liu, L. Xie, Y. Han, X. Yuan, *et al.*, "Autocaption: An approach to generate natural language description from visualization automatically," in *IEEE Pacific Visualization Symposium*, IEEE, 2020, pp. 191–195.

[91] S. Liu, W. Cui, Y. Wu, and M. Liu, "A survey on information visualization: Recent advances and challenges," *The Visual Computer*, vol. 30, no. 12, pp. 1373–1393, 2014.

[92] Z. Liu and J. Heer, "The effects of interactive latency on exploratory visual analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2122–2131, 2014.

[93] Z. Liu, N. Nersessian, and J. Stasko, "Distributed cognition as a theoretical framework for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1173–1180, 2008.

[94] J. Mackinlay, P. Hanrahan, and C. Stolte, "Show Me: Automatic presentation for visual analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1137–1144, 2007.

[95] X.-Q. Mai, J. Luo, J.-H. Wu, and Y.-J. Luo, "'Aha!' effects in a guessing riddle task: An event-related potential study," *Human Brain Mapping*, vol. 22, no. 4, pp. 261–270, 2004.

[96] R. Martinez-Maldonado, V. Echeverria, G. F. Nieto, and S. B. Shum, "From data to insights: A layered storytelling approach for multimodal learning analytics," 2020.

[97] H. Matute, F. Blanco, I. Yarritu, M. Dıaz-Lago, M. A. Vadillo, and I. Barberia, "Illusions of causality: How they bias our everyday thinking and how they could be reduced," *Frontiers in Psychology*, vol. 6, p. 888, 2015.

[98] A. McNutt, A. Crisan, and M. Correll, "Divining insights: Visual analytics through cartomancy," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–16.

[99] Z. Miao, A. Lee, and S. Roy, "LensXPlain: Visualizing and cxplaining contributing subsets for aggregate query answers," *Proceedings of the VLDB Endowment*, vol. 12, no. 12, pp. 1898–1901, 2019.

[100] L. Micallef, G. Palmas, A. Oulasvirta, and T. Weinkauf, "Towards perceptual optimization of the visual design of scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 6, pp. 1588–1599, 2017.

[101] A. V. Moere, M. Tomitsch, C. Wimmer, B. Christoph, and T. Grechenig, "Evaluating the effect of style in information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2739–2748, 2012.

[102] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," in *Biomedical Informatics*, Springer, 2014, pp. 643–674.

[103] Narrative Science. (2020). "Narrative Science." https://narrativescience.com.

[104] National Center for Education Statistics. (2020). "National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education." https://nces.ed.gov.

[105] F. Nguyen, X. Qiao, J. Heer, and J. Hullman, "Exploring the effects of aggregation choices on untrained visualization users' generalizations from data," in *Computer Graphics Forum*, Wiley Online Library.

[106] D. Norman, *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, 2013.

[107] C. North, "Toward measuring visualization insight," *IEEE Computer Graphics and Applications*, vol. 26, no. 3, pp. 6–9, 2006.

[108] J. S. Olson and W. A. Kellogg, *Ways of Knowing in HCI*. Springer, 2014, vol. 2.

[109] Oxford University Press. (2020). "insight noun — Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com." https://www.oxfordlearnersdictionaries.com/us/...

[110] A. V. Pandey, A. Manivannan, O. Nov, M. Satterthwaite, and E. Bertini, "The persuasive power of data visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2211–2220, 2014.

[111] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini, "How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2015, pp. 1469–1478.

[112] E. M. Peck, S. E. Ayuso, and O. El-Etr, "Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[113] F. Pecune, S. Murali, V. Tsai, Y. Matsuyama, and J. Cassell, "A model of social explanations for a conversational movie recommendation system," in *Proceedings of the International Conference on Human-Agent Interaction*, 2019, pp. 135–143.

[114] W. Pike, J. Bruce, B. Baddeley, D. Best, L. Franklin, R. May, D. Rice, R. Riensche, and K. Younkin, "The scalable reasoning system: Lightweight visualization for distributed analytics," *Information Visualization*, vol. 8, no. 1, pp. 71–84, 2009.

[115] C. Plaisant, J.-D. Fekete, and G. Grinstein, "Promoting insight-based evaluation of visualizations: From contest to benchmark repository," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 120–134, 2007.

[116] O. U. Press. (2021). "fact noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com." https://www.oxfordlearnersdictionaries.com/...

[117] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender system," in *Proceedings of the Fifth ACM Conference on Recommender Systems*, ACM, 2011, pp. 157–164.

[118] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 31–40, 2015.

[119] E. D. Ragan, J. R. Goodall, and A. Tung, "Evaluating how level of detail of visual history affects process memory," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2015, pp. 2711–2720.

[120] D. N. Rapp, S. R. Hinze, K. Kohlhepp, and R. A. Ryskin, "Reducing reliance on inaccurate information," *Memory & Cognition*, vol. 42, no. 1, pp. 11–26, 2014.

[121] M. Riedl. (2017). "Human-Centered Artificial Intelligence." https://medium.com/@mark_riedl/human-centered...

[122] J. Ritchie, D. Wigdor, and F. Chevalier, "A lie reveals the truth: Quasimodes for task-aligned data presentation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–13.

[123] J. M. Rohrer, "Thinking clearly about correlations and causation: Graphical causal models for observational data," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 1, pp. 27–42, 2018.

[124] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 240–249, 2015.

[125] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604–1613, 2014.

[126] P. Saraiya, C. North, and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 4, pp. 443–456, 2005.

[127] S. Sarawagi, R. Agrawal, and N. Megiddo, "Discovery-driven exploration of OLAP data cubes," in *International Conference on Extending Database Technology*, Springer, 1998, pp. 168–182.

[128] A. Sarikaya and M. Gleicher, "Scatterplots: Tasks, data, and designs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 402–412, 2017.

[129] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[130] J. Seo and B. Shneiderman, "A rank-by-feature framework for interactive exploration of multidimensional data," *Information Visualization*, vol. 4, no. 2, pp. 96–113, 2005.

[131] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao, "Calliope: Automatic visual data story generation from a spreadsheet," *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[132] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *IEEE Symposium on Visual Languages*, IEEE, 1996, pp. 336–343.

[133] T. Siddiqui, A. Kim, J. Lee, K. Karahalios, and A. Parameswaran, "Effortless data exploration with zenvisage: An expressive and interactive visual analytics system," *Proceedings of the VLDB Endowment*, vol. 10, no. 4, 2016.

[134] R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ACM, 2002, pp. 830–831.

[135] Sperling's Best Places. (2020). "2020 Compare Climate Weather: Seattle, WA vs Atlanta, GA." `https://www.bestplaces.net/climate/...`

[136] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko, "Augmenting visualizations with interactive data facts to facilitate interpretation and communication," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 672–681, 2018.

[137] J. Stasko, "Value-driven evaluation of visualizations," in *Proceedings of the 2014 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, 2014, pp. 46–53.

[138] ——, (2019). "What's an Insight? – As I See It." https://jts3blog.wordpress.com/2018/02/22/whats-an...

[139] J. H. Stock and M. W. Watson, *Introduction to Econometrics*. 2015.

[140] R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data & Knowledge Engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.

[141] L. A. Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge university press, 1987.

[142] Tableau. (2017). "MEDIA ALERT: Tableau Public Surpasses One Million Visualizations." https://www.tableau.com/about/press-releases...

[143] ——, (2019). "Explain Data Internals: Automated Bayesian Modeling | Tableau Conference 2019." https://tc19.tableau.com/learn/sessions...

[144] ——, (2019). "Inspect a View using Explain Data – Tableau." https://help.tableau.com/current/pro/desktop/en-us/...

[145] ——, (2020). "Explain Data | Tableau Software." https://www.tableau.com/products/new-features/explain...

[146] ——, (2020). "Tableau." https://www.tableau.com.

[147] B. Tang, S. Han, M. L. Yiu, R. Ding, and D. Zhang, "Extracting top-k insights from multi-dimensional data," in *Proceedings of the 2017 International Conference on Management of Data*, 2017, pp. 1509–1524.

[148] H. Thomson. (2018). "Aha! What Happens in Your Brain When You Have a Lightbulb Moment | New Scientist." https://www.newscientist.com/...

[149] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA, 1977, vol. 2.

[150] United States Census Bureau. (2020). "Census Bureau." https://www.census.gov.

[151] A. C. Valdez, M. Ziefle, and M. Sedlmair, "Priming and anchoring effects in visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 584–594, 2017.

[152] M. Vartak, S. Rahman, S. Madden, A. Parameswaran, and N. Polyzotis, "SeeDB: Efficient data-driven visualization recommendations to support visual analytics," *Proceedings of the VLDB Endowment*, vol. 8, no. 13, pp. 2182–2193, 2015.

[153] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon, "Many eyes: A site for visualization at internet scale," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1121–1128, 2007.

[154] ——, "Many eyes: A site for visualization at internet scale," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1121–1128, 2007.

[155] J. Wang and K. Mueller, "The visual causality analyst: An interactive interface for causal reasoning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 230–239, 2015.

[156] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, and D. Zhang, "DataShot: Automatic generation of fact sheets from tabular data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 895–905, 2019.

[157] N. Whitman. (2013). "True Facts and False Facts : Behind the Dictionary : Thinkmap Visual Thesaurus."
`https://www.visualthesaurus.com/cm/dictionary/...`

[158] L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *IEEE Symposium on Information Visualization*, IEEE, 2005, pp. 157–164.

[159] W. Willett, J. Heer, and M. Agrawala, "Scented widgets: Improving navigation cues with embedded visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1129–1136, 2007.

[160] W. Willett, J. Heer, J. Hellerstein, and M. Agrawala, "Commentspace: Structured support for collaborative visual analysis," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 3131–3140.

[161] G. Wills and L. Wilkinson, "AutoVis: Automatic visualization," *Information Visualization*, vol. 9, no. 1, pp. 47–69, 2010.

[162] M. S. Wogalter, "Communication-human information processing (c-hip) model," *Handbook of warnings*, pp. 51–61, 2006.

[163] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager: Exploratory analysis via faceted browsing of visualization recommendations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 649–658, 2015.

[164] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager 2: Augmenting visual analysis with partial view specifications," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2017, pp. 2648–2659.

[165] J. Xiao, J. Stasko, and R. Catrambone, "An empirical study of the effect of agent competence on user performance and perception," in *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems*, IEEE, 2004, pp. 178–185.

[166] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri, "Illusion of causality in visualized data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 853–862, 2019.

[167] C. Xiong, L. van Weelden, and S. Franconeri, "The curse of knowledge in visual data communication," *IEEE Transactions on Visualization and Computer Graphics*, 2019.

[168] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko, "Understanding and characterizing insights: How do people gain insights using information visualization?" In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, 2008, pp. 1–6.

[169] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska, "Investigating the effect of the multiple comparisons problem in visual analysis," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[170] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. Keim, "Visual analytics for the big data era—a comparative review of state-of-the-art commercial systems," in *IEEE Conference on Visual Analytics Science and Technology*, IEEE, 2012, pp. 173–182.

[171] M. Zhao, H. Qu, and M. Sedlmair, "Neighborhood perception in bar charts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–12.

[172] Z. Zhao, L. De Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska, "Controlling false discoveries during interactive data exploration," in *Proceedings of the ACM International Conference on Management of Data*, 2017, pp. 527–540.